Some Contributions to Audio Source Separation and Diarisation of Multichannel Convolutive Mixtures

Dionyssos Kounades-Bastian

Communauté
UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ DE GRENOBLE

Spécialité : **Mathématiques et Informatique**

Arrêté ministériel :

Présentée par

Dionyssos Kounadis-Bastian

Thèse dirigée par **Radu Horaud**
et codirigée par **Laurent Girin**

préparée au sein **de l'Université de Grenoble et de INRIA Grenoble Rhône-Alpes**
et de **L'École Doctorale de Mathématiques, Sciences et Technologies de l'Information, Informatique**

Quelques Contributions pour la Séparation et la Journalisation de Sources Audio

dans des Mélanges Multicanaux Convolutifs

Thèse soutenue publiquement le **24 Février 2017**,
devant le jury composé de :

Pr. Jérôme Mars
Grenoble INP, Président
Pr. Emmanuel Vincent
INRIA Nancy Grand Est, Rapporteur
Pr. Cédric Févotte
Institut de Recherche en Informatique de Toulouse, Rapporteur
Pr. Roland Badeau
Institut Mines-Télécom, Tlécom ParisTech, CNRS LTCI, Examinateur
Pr. Tuomas Virtanen
Tampere University of Technology, Examinateur
Dr. Xavier Alameda-Pineda
University of Trento, Examinateur
Pr. Laurent Girin
Grenoble INP, Co-Directeur de thèse
Pr. Radu Horaud
INRIA Grenoble Rhône-Alpes, Directeur de thèse

Dionyssos Kounadis-Bastian

Some Contributions to

Audio Source Separation and Diarisation

of Multichannel Convolutive Mixtures

Grenoble, December 21, 2016

Abstract

In this thesis we address the problem of *multichannel audio source separation* (MASS) for underdetermined convolutive mixtures through probabilistic modeling. We focus on three aspects of the problem and make three contributions. Firstly, inspired from the empirically well validated representation of an audio signal, that is know as *local Gaussian signal model* (LGM) with *non-negative matrix factorization* (NMF), we propose a Bayesian extension to this, that overcomes some of the limitations of the NMF. We incorporate this representation in a MASS framework and compare it with the state of the art in MASS, yielding promising results. Secondly, we study how to separate mixtures of moving sources and/or of moving microphones. Movements make the *acoustic path* between sources and microphones become *time-varying*. Addressing time-varying audio mixtures appears is not so popular in the MASS literature. Thus, we begin from a state of the art LGM-with-NMF method designed for separating time-invariant audio mixtures and propose an extension that uses a Kalman smoother to track the acoustic path across time. The proposed method is benchmarked against a block-wise adaptation of that state of the art (ran on time segments), and delivers competitive results on both simulated and real-world mixtures. Lastly, we investigate the link between MASS and the task of audio diarisation. Audio diarisation is the detection of the time intervals where each speaker/source is active or silent. Most state of the art MASS methods consider the sources to emit continuously; A hypothesis that can result in spurious signal estimates for a source, in intervals where that source was silent. Our aim is that diarisation can aid MASS by indicating the emitting sources at each time frame. To that extent we design a joint framework for simultaneous diarisation and MASS, that incorporates a hidden Markov model (HMM) to track the temporal activity of the sources, within a state of the art LGM-with-NMF MASS framework. We compare the proposed method with the state of the art in MASS and audio diarisation tasks. We obtain performances comparable, with the state of the art, in terms of separation while winning in terms of diarisation.

Résumé

Dans cette thèse nous abordons le problème de la séparation de sources audio dans des mélanges convolutifs multicanaux et sous-déterminés, en utilisant une modélisation probabiliste. Nous nous concentrons sur trois aspects, et nous apportons trois contributions. D'abord, nous nous inspirons du modèle Gaussien local par factorisation en matrices non-négatives (LGM-with-NMF), qui est un modéle empiriquement validé pour représenter un signal audio. Nous proposons une extension Bayésienne de ce modèle, qui permet de surpasser certaines limitations du modèle NMF. Nous incorporons cette representation dans un cadre de separation audio multicanaux, et le comparons avec l'état de l'art sur des tâches de separation. Nous obtenons des résultats prometteurs. Deuxièment, nous étudions comment séparer des mélanges audio de sources et/ou des capteurs en movement. Ces déplacements rendent le chemin acoustique entre les sources et les microphones variant en cours du temps. L'adressage des mélanges convolutifs variant au cours du temps est peu exploré dans la littérature. Ainsi, nous partons d'une methode de l'état de l'art développée pour la séparation de melanges invariant (sources et microphones statiques) et utilisant LGM-with-NMF. Nous proposons a ceci une extension qui utilise un filtre de Kalman pour suivre le chemin acoustique au cours du temps. La technique proposée est comparée à une adaptation block-par-block d'une technique de l'état de l'art appliquée sur des intervalles de temps, et a donné des résultats exceptionels sur les melanges simulés et les melanges du monde réel. Enfin, nous investiguons les similitudes entre la separation et la journalisation audio. La journalisation est le probléme de détection des intervalles auxquels chaque locuteur/source est émettant. La plupart des méthodes de séparation supposent toutes les sources émettent continuellment. Cette hypothèse peut donner lieu à de fausses estimations durant les intervalles au cours desquels cette source n'a pas émis. Notre objectif est que la journalisation puisse aider à résoudre le separation, en indiquant les sources qui eméttent à chaque intervalle de temps. Dans cette mesure, nous concevons une cadre commun pour traiter simultanément la journalisation et la separation du mélange audio. Ce cadre incorpore un modèle de Markov caché pour suivre les activités des sources au sein d'une technique de séparation LGM-with-NMF. Nous comparons l'algorithme proposé à l'état de l'art sur des tâches de separation et de journalisation. Nous obtenons des performances comparables avec l'etat de l'art pour la separation, et supériures pour la journalisation.

Zussamenfassung

In dieser Doktorarbeit beschäftigen wir uns mit dem Problem der Trennung von Schallquellen, der mehrkanalig und unterbestimmten Faltungsmischungen mittels probabilisticher Modellierung. Wir konzentrieren uns auf drei Aspekte des Problems und leisten drei Beiträge dazu. Erstens, inspiriert von dem empirisch gut bestätigtem Ansatz für Signaldarstellung, der als das lokale Gaußische Modell mit nichtnegativer Matrixfaktorisierung bekannt ist; Schlagen wir eine Bayesianische Erweiterung vor, die einige Begrenzungen der nichtnegativer Matrixfaktorisierung aufhebt. Wir setzen diese Darstellung in einen mehrkanaligen Trennungsrahmen und vergleichen es mit dem Stand der Technik zur Schallquellentrennung. Wir erhalten vielversprechende Resultate. Zweitens, untersuchen wir die Weise auf welche man Mischungen von bewegenden Schallquellen und mobilen Mikrofonen trennen kann. Solche Bewegungen gestalten den akustischen Pfaden zwischen den Schallquellen und den Mikrofonen als zeitvariabel. Soweit wir wissen, die Algorihtmen für Trennung der zeitvariabelnden Faltungsmischungen sind selten in der Fachliteratur. Deswegen, beginnen wir mit einer Stand der Technik Trennugsmethode die für zeitlich invariablen Faltungsmischungen erbaut wurde. In die Methode stecken wir einen Kalman Filter der die Laufbahn des Schalles aufspürt. Die vorgeschlagene Methode wird verglichen mit einer blockweise Anpassung aus Zeitsegmenten der Stand der Technik Methode. Unsere Methode liefert hervorragende Resultate, sowohl bei den Mischungen aus der simulierten Realität als auch aus der Wirklichkeit. Letztens, wir untersuchen die Beziehung zwischen der Schallquellentrennung und der Diarisierung. Diarisierung ist die Anmerkung von Zeitabschnitten wo jeder Sprecher/Quelle Schalllos ist. Die meisten Schallquellentrennungsmethoden betrachten die Quellen als unaufhörlich emittierend. Diese Annahme könnte fadenscheinige Signalschätzungen liefern im Laufe der Zeitspannen derer die Quelle nicht emittierte. Wir wollen der Trennung durch die Diarisierung helfen, mittels der Anzeige der emittierenden Schallquellen. Um die gleichzeitige Diarisierung und Trennung der Faltungsmischungen zu erreichen, setzen wir ein Hidden Markov Model ein für die nachführung der Aussendung der Schallquellen in eine Stand-der-Technik Trennungsmethode. Wir vergleichen den vorgeschlagenen Algorithmus mit Stand der Technik Methoden bei Aufgaben der Trennung und der Diarisierung. Das ergibt ähnliche Resultate bezüglich der Trennung und überragende hinsichtlich der Diarisierung.

Περὶληψη

Σε αυτή τη διατριβή ασχολούμαστε με το πρόβλημα του διαχωρισμού ηχητικών πηγών επεξεργαζόμενοι τα συνελικτικά μίγματα σημάτων αυτών των πηγών που λαμβάνουμε μέσα από πολλά κανάλια, με χρήση μοντέλων πιθανοτήτων. Εστιάζουμε σε τρείς πτυχές του προβλήματος και προτείνουμε τρείς επεκτάσεις. Πρώτον. Εμπνευσμένοι από την πειραματικά επαληθευμένη αναπαράσταση ενός ηχητικού σήματος που ονομάζεται τοπική Γκαουσιανή μοντελοποίηση με παραγοντοποίηση σε θετικά μητρώα, προτείνουμε μια προσέγγιση κατά Bayes που ξεπερνά αδυναμίες της συμβατικής παραγοντοποίησης σε θετικά μητρώα. Ενσωματώνουμε την προτεινόμενη αναπαράσταση σε ένα αλγοριθμικό πλαίσιο διαχωρισμού, μίγματος πολυκαναλικού σήματος και την συγκρίνουμε με αλγοριθμικές μεθόδους αιχμής αποκομίζοντας θετικά αποτελέσματα. Δεύτερον. Μελετώντας πως μπορούμε να διαχωρίσουμε ηχιτικά μίγματα που μπορεί να προέρχονται από κινούμενες πηγές, αλλά και να λαμβάνονται και από κινούμενα μικρόφωνα; Παρατηρήσαμε ότι η κινήση επηρεάζει την ακουστική ζεύξη χρονικά μεταβαλλόμενη. Η επεξεργασία τέτοιων μιγμάτων δεν απαντάται συχνά στην επιστημονική βιβλιογραφία. Υπό το πρίσμα αυτής της παρατήρησης χτήζουμε πάνω με μια μέθοδο διαχωρισμού ακίνητων πηγών στην οποία εισάγουμε ένα φίλτρο Kalman για την ιχνηλάτηση της ακουστικής ζεύξης στο χρόνο. Συγκρίνουμε την προτεινόμενη τεχνική ενάντίων της τεχνική αιχμής εφαρμοσμένη ανα χρονικά διαστήματα, σε προβλήματα διαχωρισμό συνθετικών χρονικλα μεταβαλλόμενων μίγματων, αλλά και πραγματικών ηχογραφήσεων κινουμένων ομιλητών. Η σύγκριση αποδίδει σημαντικά πειραματικά αποτελέσματα υπέρ της τεχνικής μας. Τρίτον. Ως επί το πλείστον οι υπάρχουσες τεχνικές διαχωρισμού θεωρούν ότι οι πηγές εκπέμπουν αδιαλείπτως. Αυτη η υπόθεση μπορεί να δώσει εσφαλμένες εκτιμήσεις σε διαστήματα που οι πηγές σιωπούν. Ετσι λοιπόν, εξερευνούμε τις ομοιότητες μεταξύ διαχωρισμού και Καταγραφής Ημερολογίου Εκπομπής και κατασκευάζουμε μια τεχνική ταυτόχρονης επίλυσης των δύο προβλημάτων, αφού εισάγουμε ένα λανθάνων μακροβιανό μοντελο να ιχνηλάτεί την εκπομπή κάθε πηγής, μέσα σε μια μέθοδο διαχωρισμού. Τα αποτελέσματα μας κάνουν να αισιοδοξούμε.

ACKNOWLEDGMENT

Tout d'abord je dois un grand merci à Radu qui m'a proposé de commencer cette thèse trois ans plus tôt. Son style en tant que superviseur mais aussi sa confiance à moi tout au long de cette chemin m'agissent comme des leçons de professionnalisme. Je suis fier d'avoir travaillé avec lui.

Je dois aussi un égal remerciement à Laurent pour les innombrables heures qu'il a passées à scorriger mes rapports burlesques et sa passion sans limites pour la recherche.

Xavi para ese informe de ocho páginas sobre el modelo probabilístico de siete latentes, Que en esencia contenía la totalidad de esta tesis y mucho más..

Sharon for all the courses on audio signal processing.

Δάσκαλε κ. Ψαράκη γιατι μου είπατε "*Πίστεψε στην θεωρία και βάλε τα πειράματα να επαληθεύσουν*!

Γιώργο για τις συμβουλές σου και το τράβηγμα να περάσω μία βόλτα απο Ευρώπη!

Ότι και να πώ είναι λίγο για τούς φίλους μου, Εύαγγελο και Μαρία Τερέζα, και το ταξίδι αυτά τα 27 *χρόνια*.

The team, for all the random moments, the ideas and their inspiration: Antoine, Xiaofei, Yutong, Soraya, Benoit, Stephan, Quentin, Gulliaume, Bastien, Pablo, Rémi, Fabien, Pier, Jordi, and all people i forgot to mention..

Nathalie pour tout ton soutien.. tu avais tout préparé pour nous tous .. Je n'avais qu'à monter à bord de l'avion!

Δέν υπάρχει περίπτωση να ξεχάσω την παρέα: *Δημήτρη* well played..

Vincent, '*Αντρεα*, '*Αντρια*, Israel, Daan, Hong, pour mon introduction à la société,

Rambo and *Χριστουλάκι* for our freddocinos in NS square and your perpetual passion to overtake the 0.7,

Τσόντας και Σμούρι για όλες τις στιγμές ασοβαρότητας,

Ιωάννα σε μία στιγμή μου είπες "'*Οπου και αν βρεθείς άφησε μια ωραία γνώμη*".

A great thank you to my examining committee for thoroughly evaluating this manuscript.

Illusory moments of bliss

Contents

LIST OF FIGURES

LIST OF TABLES

NOTATION

NOMENCLATURE

$\mathbf{A}_f$	Boldface upper case letters denote matrices.
$\mathbf{a}_f$	Boldface lower case letters denote column vectors.
$\mathbf{A}_{f1:L}$	A compact way to denote a set, here $\{\mathbf{A}_{f\ell}\}_{\ell=1}^{L}$.
$a_{j,f}$	The j^{th} entry of $\mathbf{a}_f$.
$A_{ij,f}$	The $(i,j)^{\text{th}}$ entry of $\mathbf{A}_f$.
θ	The set of all parameters to be estimated, of a generative model.
$\mathcal{H}$	The set of all hidden variables of a generative model.

SIZES

$f \in [1, F]$	Number of frequency bins F, and frequency index f.
$\ell \in [1, L]$	Number of STFT (time) frames L, and frame index ℓ.
$i \in [1, I]$	Number of microphones I, and microphone index i.
$j \in [1, J]$	Number of sources J, and source index j.
$k \in [1, K]$	Total Number of LGcM components K, and index k.
$n \in [1, N]$	Number of diarisation states N, and state index n.

FUNCTIONS AND OPERATORS

$\mathbf{A}^\top$	Matrix transpose (without conjugation).
$\mathbf{A}^{\mathrm{H}}$	Hermitian transpose of a matrix.
$\mathbf{A}^{-1}$	Inverse of a matrix.
$\det(\mathbf{A})$	Determinant of a matrix.
$\mathrm{tr}\{\mathbf{A}\}$	Trace of a matrix.
$\mathrm{vec}\,(\mathbf{A})$	The column vector made by concatenating all columns of $\mathbf{A}$ in a single vector.
$\otimes$	Kronecker (matrix) product.
$\mathbf{I}_J$	Identity matrix of dimension J.
$\mathrm{diag}_J\,(a_j)$	$J \times J$ diagonal matrix with entries $\{a_j\}_{j=1}^J$.
$\lvert a\rvert^2 = aa^{\mathrm{H}}$	Squared modulus of complex number.
$\mathfrak{Re}\{a\}$	Real part of complex number.
$\pi_n \overset{n}{\propto} y_n$	Normalisation with $\pi_n = y_n \big/ \sum_{r=1}^{N} y_r$.
$\log(x)$	Natural logarithm, at the base $\exp(1)$.
$\Gamma(x)$	Gamma function, with $x \in \mathbb{R}_+$.
$\psi(x)$	Digamma function, with $x \in \mathbb{R}_+$.
$\mathbb{E}_{p(x)}[f(x)]$	The expected value of $f(x)$ with respect to probability distribution $p(x)$.
$\mathcal{N}_c(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\exp\left(-(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{H}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\det\left(\pi\boldsymbol{\Sigma}\right)}$	Circularly-symmetric complex normal distribution [Neeser 93] for $\mathbf{x} \in \mathbb{C}^I$, mean vector $\boldsymbol{\mu} \in \mathbb{C}^I$, covariance matrix $\boldsymbol{\Sigma} \in \mathbb{C}^{I\times I}$.
$\mathcal{IG}\,(x;\gamma,\delta) = \frac{(\delta)^\gamma}{\Gamma(\gamma)} x^{-(\gamma+1)} \exp\left(-\frac{\delta}{x}\right)$	Inverse Gamma distribution [Witkovsky 01] for non negative $x \in \mathbb{R}_+$, with parameters: shape $\gamma \in \mathbb{R}_+$, and scale $\delta \in \mathbb{R}_+$.

CHAPTER 1

INTRODUCTION

In robot audition it is a key challenge to discriminate the sound sources that make up the recorded audio signal at a microphone array, that is called the mixture signal. Audio source separation is the scientific field encompassing techniques that recover the sound source signals from their mixture signals. Audio source separation is nowadays a key ingredient of speech recognition and machine translation. Its theoretical background extends beyond audio processing on various scientific fields, such as biomedical imaging and image processing. In the past fourty years the effervescent research conducted on this field established probabilistic modeling as one of the prominent directions to address source separation. In this thesis we investigate the *source separation from multichannel audio mixtures* (MASS). By taking a technical look on latent structures present in natural sound signals and their generative process we design probabilistic methods aiming to separate and diarise multichannel audio mixtures. Our major focus is underdetermined mixtures (fewer microphones than sources) with moving sources. In this introductory chapter we give an overview of MASS and present the main scientific roads that have been taken to address it. We present the Local Gaussian Model (LGM) for sound signals, as it is a core ingredient of all designs of this thesis. Finally, we summarise our contributions and plot the organisation of this manuscript.

1.1 INSPIRATION

The majority of everyday sound scenes involve several sound sources that emit simultaneously. Speech communication is obscured by background talkers and environmental sounds interfering to the conversation. When facing such situations, humans are at ease on concentrating at any of the individual sound sources [Cherry 53, Wang 07]. In audio source separation we want to design algorithms to recover the original sound source signals, from recordings of the overall sound scene, that are known as mixture signals. The general source separation problem is equivalent to answer the question: "*Given the sum of*

two numbers could you recover the individual summands?" Such a general problem with no additional information has no unique solution; however in MASS we are opted with a large amount of extra information. Natural sounds reveal sparsity in some domains, the most well being the Short Time Fourier Transform (STFT) domain [Portnoff 80]. Recent studies in music audio processing show that sound signals have structures and can be well represented in a *dictionary-activation* basis [Benaroya 03, Févotte 09] with few parameters. Those characteristics of sound enable the design of MASS algorithms that can deliver surprising results.

1.2 THE AUDIO SOURCE SEPARATION PROBLEM

In this thesis we are interested in indoor recordings. Commonplace indoor environments introduce adverse effects on the recorded mixture signal, such as reverberation. The presence of reverberation makes the MASS problem somewhat easier to solve in the STFT domain than in the time-domain.

1.2.1 AUDIO MIXTURES IN THE TIME DOMAIN

If we assume the existence of J sound sources and denote with $y_{ij}(t)$ the contribution of j-th source to microphone i we can write the signal $x_i(t)$ recorded at microphone i as:

$$x_i(t) = \sum_{j=1}^{J} y_{ij}(t) + b_i(t), \tag{1.1}$$

with $b_i(t)$ a noise signal, for example from the sensors. In real world scenarios there may be more sources than microphones ($J > I$). Such scenario is called *underdetermined mixing*, in contrast to the *(over)determined* mixing where $J \leq I$. The overdetermined MASS case has been long studied and nowadays overdetermined MASS methods can provide good source separation performance [Gannot 17]. In this thesis we are interested in underdetermined mixtures of indoor recordings.

The major effect of indoor recordings is reverberation. That is the fact that the microphones capture not only the direct sound coming from the sources, but also its reflections from the walls and the surfaces of other objects in the scene. Therefore $y_{ij}(t)$ is the sum of all reflections of source j as they arrive at microphone i [Duong 10, Sturmel 12]. The $y_{ij}(t)$ is known as the *source image* signal and is defined as the convolution (denoted $'*'$) of the source signal $s_j(t)$ with an impulse response signal $a_{ij}(t)$:

$$y_{ij}(t) = a_{ij}(t) * s_j(t). \tag{1.2}$$

The impulse response $a_{ij}(t)$, that is called the *mixing filter* and is encoding the acoustical effects induced by the environment, such as reverberation. Eq. (1.1) and (1.2) define a

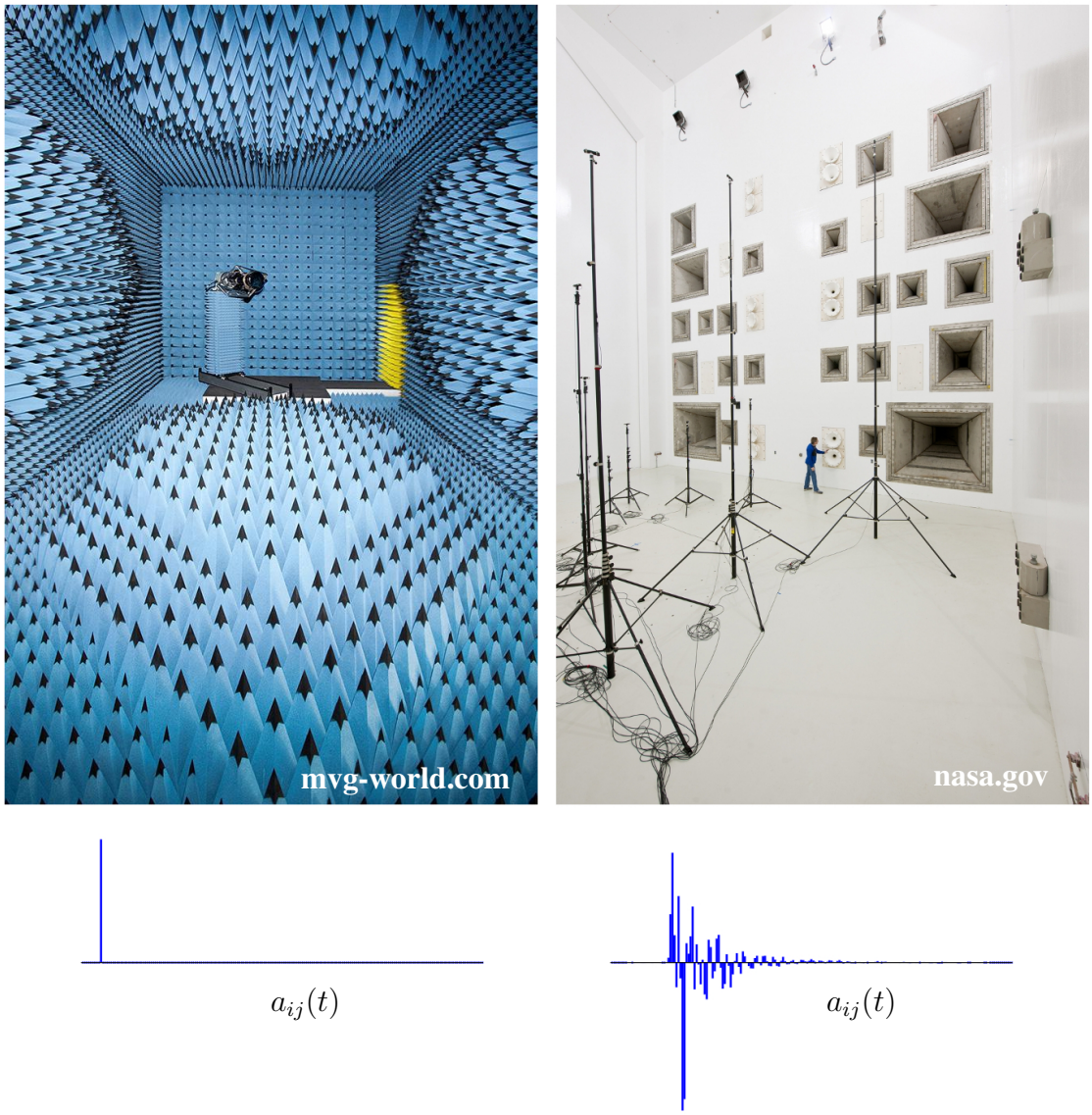

Figure 1.1: Filter responses: of an anechoic (low echo) room (**Left**) with special walls to reduce reverberation, and a chamber with high reverberation (**Right**).

convolutive mixture:

$$x_i(t) = \sum_{j=1}^{J} a_{ij}(t) * s_j(t) + b_i(t). \tag{1.3}$$

The MASS problem can be stated as the recovery of the source signals $\{s_j(t)\}_{j=1}^{J}$ from the mixture signals $\{x_i(t)\}_{i=1}^{I}$. Notice here that the $\{a_{ij}(t)\}_{i,j=1}^{I,J}$ are generally unknown and have to be estimated as well.

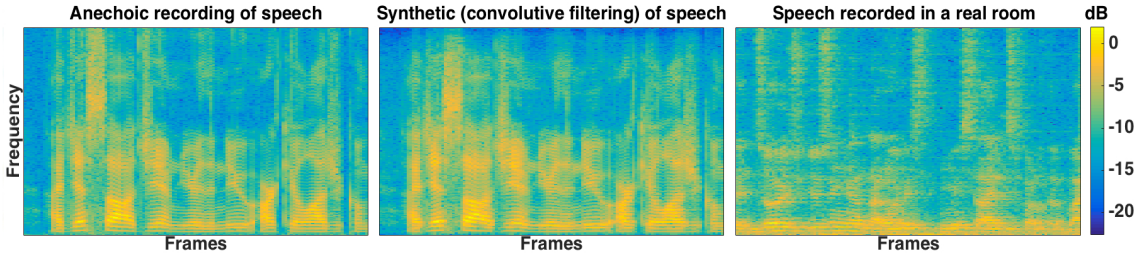

Figure 1.2: Spectrograms of convolutive (simulated) and real recordings of speech in indoor environments. The presence of a high level of sparsity and harmonic structure is apparent, although less visible on the real recordings.

1.2.2 Audio Mixtures in the Time-Frequency Domain

The presence of convolution in (1.3) complicates the design of time-domain MASS methods. STFT has prevailed as way to transform the MASS task in a time-frequency representation, opting for a simpler solution.

Applying the STFT to the mixture signal of the i-th microphone yields a set of complex-valued coefficients $\{x_{i,f\ell}\}_{f,\ell=1}^{F,L}$ for the F frequency bins and the L time-frames. The mixture $\mathbf{x}_{f\ell} \in \mathbf{C}^I$ is approximated in the STFT with:

$$\mathbf{x}_{f\ell} = \mathbf{A}_f \mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}, \tag{1.4}$$

with $\mathbf{A}_f \in \mathbb{C}^{I\times J}$ the mixing matrix, $\mathbf{s}_{f\ell} \in \mathbb{C}^J$ the vector of source coefficients and $\mathbf{b}_{f\ell}$ some residual noise. Now the MASS task becomes the recovery of $\{\mathbf{s}_{f\ell}\}_{f,\ell=1}^{F,L}$ and of $\{\mathbf{A}_f\}_{f=1}^{F}$. Eq.(1.4) is known as the *narrowband assumption* [Parra 00, Gannot 01, Ozerov 12] and is valid for environments with low reverberation, becoming less appropriate as reverberation increases. Eq.(1.4) is popular due to its simplicity; it may be overcome by directly recovering the source-images [Duong 10, Arberet 10], or by designing a detailed reverberation model [Leglaive 16], or by working in the time domain [Kowalski 10].

1.3 Literature Overview

Today, the state of the art in MASS for convolutive mixtures is vast. A comprehensive survey can be found in [Gannot 17]. We split the state of the art in three non-exclusive categories.

Methods using localisation information A large family of MASS techniques use information about the underlying spatial location of the sources. *Computational Auditory Scene Analysis* (CASA) intent to emulate the human auditory scene formation process [Blauert 97, Wang 07]. CASA methods are based on the apparent sparsity of speech in

time-frequency (TF) representations. In the core of CASA systems the mixture signal is transformed in a TF representation and the TF points are clustered in groups associated with a single source. Popular criteria used for clustering include interchannel time or intensity differences of TF points [Yilmaz 04, Araki 07]. TF points with similar such differences must have been generated from the same spatial location. At end of clustering the estimated TF representation for a source is populated with TF points of the mixture that are clustered on that respective source. Empty TF ponts are filled with zero and the inverse TF transform is applied to provide the time-domain estimate for that source. The limitation of CASA methods is the assumption that a TF point contains information from a single source. In real world recordings, where reverberation is substantial, CASA methods can be limited [Araki 03].

Beamforming MASS methods [Hioka 13] enhance the sound coming from a specific location in the room and attain separation by enhancing the signals from the locations of the sources.

Independent Component Analysis (ICA) ICA is a method for separating a multichannel signal (the mixture) into additive components (the sources) [Hyvärinen 01]. The principle of ICA methods is to assume the underlying source signals as statistically independent [Cardoso 98]. Because then various criteria for extracting components that are *more independent than the mixture* can be used to recover the sources signals [Cardoso 97]. For MASS, ICA is applied independently at each frequency f with (1.4) [Sawada 04, Sawada 07]. Because the source signals will be recovered with a different order at each frequency a realignment to make them correspond to the same source over all frequencies is needed. This alignment is done in a second step by exploiting relations between mixing matrices from different frequencies. ICA methods may not be applied to underdetermined mixtures due to the requirement of an invertible $\mathbf{A}_f$.

Probabilistic Inference for Source Separation Insufficiency of observed data (when we have fewer microphones than sources) places a strong barrier on the recovery of high quality separated signals. For this reason, methods for separating underdetermined mixtures use prior knowledge about the sound production process. The knowledge typically concerns the structure of the underlying source signals and the generating process of the mixture. Such methods are referred to as *model based* MASS [Mandel 10]. Model based methods rely on generative models for the source signals [Vincent 10] and/or the mixture [Dorfan 15]. Typically the source signals are considered as hidden random variables and prior *probability distribution functions* (PDF) are placed on them [Févotte 09]. The separated source signals are obtained through statistical estimation criteria, such as *maximum likelihood* (ML) or *maximum posterior* (MAP) estimation [Vincent 10]. A practical generative model includes numerous additional *model parameters* to be estimated. In generative models it is now classical to use an *Expectation Maximization* (EM) algorithm for inference of the hidden variables and learning of the model parameters [Ozerov 12]. One of the popular frameworks for audio signal modeling in the STFT domain, that will be an important ingredient of this thesis, is the local Gaussian source model (LGM) [Benaroya 03].

1.4 PROBABILISTIC INFERENCE FOR SOURCE SEPARATION

Various works on model based MASS in the STFT domain consider the source coefficients as hidden random variables with prior distributions, for example [Ozerov 10, Ozerov 12, Arberet 10, Leglaive 16]. Such methods use MAP to estimate the source coefficients and apply the inverse STFT to them so as to obtain the time domain source signal estimates. To apply MAP to a generative model the *posterior probability distribution* of its hidden variables must be infered. In practical generative models, besides the hidden variables, there are various model parameters that have to be estimated as well. For that an Expectation Maximization (EM) algorithm [Bishop 06] is used to infer the hidden variables and estimate the model parameters.

1.4.1 THE EXPECTATION MAXIMIZATION ALGORITHM

EM is an iterative optimization algorithm that finds ML estimates for the model parameters of a generative model, in the presence of hidden variables. EM alternates between evaluation of the posterior probability distribution function (PDF) of the hidden variables, called the E step, and maximisation, with respect to the model parameters, of the *expected complete data log-likelihood* (ECDLL) function, called the M step [Bishop 06].

A generative model is specified by a set of hidden variables $\mathcal{H}$, a set of observed data (for STFT domain MASS the coefficients of the mixture $\mathbf{x}_{1:F1:L}$), a set of model parameters θ, and a *complete data distribution* $p\left(\mathbf{x}_{1:F1:L}, \mathcal{H}; \theta\right)$ typically in parametric form. To design an EM algorithm we first compute the posterior probability distribution:

$$p\left(\mathcal{H}|\mathbf{x}_{1:F1:L}\right). \tag{1.5}$$

This makes the E step.[1] Then, we calculate the ECDLL denoted $\mathcal{L}(\theta)$ and defined with:[2]

$$\mathcal{L}(\theta) = \mathbb{E}_{p(\mathcal{H}|\mathbf{x}_{1:F1:L})}\left[\log p\left(\mathcal{H}, \mathbf{x}_{1:F1:L}; \theta\right)\right]. \tag{1.6}$$

Maximising $\mathcal{L}(\theta)$ with respect to θ results in the updated values for θ. This makes the M step. The E and M steps are iterated[3] until a convergence criterion is met.

Variational EM In complicated generative models the posterior distribution may not be expressible in terms of standard distributions (due to intractable integrals). In such cases, the posterior distribution must be approximated. Various ways exist to approximate it, see for example [Bishop 06, Smidl 06]. In this thesis we use the so called *variational approximation* [Jordan 99, Bishop 06]. In the variational the set of hidden variables $\mathcal{H}$ is partitioned to P subsets $\mathcal{H} = \{\mathcal{H}_p\}_{p=1}^{P}$. Then the posterior distribution[4], denoted $q(\mathcal{H})$ is

[1] For clarity we omit writing θ in $p\left(\mathcal{H}|\mathbf{x}_{1:F1:L}\right)$.

[2] To compute $\mathcal{L}(\theta)$, $p(\mathcal{H}|\mathbf{x}_{1:F1:L})$ is provided by the E step and is fixed.

[3] In the next E step, the updated values of θ are used to compute $p(\mathcal{H}|\mathbf{x}_{1:F1:L})$.

[4] It is the approximate posterior distribution, but we abuse the language and refer to it as posterior.

assumed to factorise over the posterior distribution of the P subsets:

$$p(\mathcal{H}|\mathbf{x}_{1:F1:L}) \approx q(\mathcal{H}) = \prod_{p=1}^{P} q(\mathcal{H}_p). \tag{1.7}$$

The posterior distribution $q(\mathcal{H}_p)$ of a subset $\mathcal{H}_p$ is then computed with [Bishop 06]:[5]

$$q(\mathcal{H}_p) \propto \exp\left(\mathbb{E}_{q(\mathcal{H}/\mathcal{H}_p)}\left[\log p(\mathcal{H}, \mathbf{x}_{1:F1:L}; \theta)\right]\right), \tag{1.8}$$

with $q(\mathcal{H}/\mathcal{H}_p)$ being the product of all $q(\mathcal{H}_p)$ of all other subsets, except of p. Easily, the full posterior $q(\mathcal{H})$ is computed with (1.7). Hence, if we use $q(\mathcal{H})$ in place of (1.5) we have an E step. As for the M step, one has to define $\mathcal{L}(\theta)$ using $q(\mathcal{H})$ in place of $p(\mathcal{H}|\mathbf{x}_{1:F1:L})$ with:

$$\mathcal{L}(\theta) = \mathbb{E}_{q(\mathcal{H})}\left[\log p\left(\mathcal{H}, \mathbf{x}_{1:F1:L}; \theta\right)\right]. \tag{1.9}$$

This makes the variational EM (vEM). In summary, at the E step we compute $q(\mathcal{H})$ and at the M step we update θ by optimising (1.9).

1.4.2 Local Gaussian Source Models

For the past decade the statistical modeling of audio signals in the time-frequency domain has been extensively investigated. The LGM is a prominent example of such modeling and has become popular in MASS as a parsimonious representation for the source signals. In LGM the STFT coefficients of the source are assigned with a prior PDF that is a Gaussian PDF whose support are the complex numbers [Ephraim 84].

To reduce the number of parameters to be estimated [Benaroya 03] introduced a *non-negative matrix factorisation* (NMF) scheme on the variance of that prior PDF. The resulting model is known as the *Local Gaussian composite Model* (LGcM) with NMF:

$$p(s_{j,f\ell}) = \mathcal{N}_c\left(s_{j,f\ell}; 0, \sum_{k\in\mathcal{K}_j} w_{fk}h_{k\ell}\right), \tag{1.10}$$

with $w_{fk}, h_{k\ell} \in \mathbb{R}_+$ parameters to be estimated, and $\mathcal{K}_j$ a subset indicating the indexes (of the factors $w_{fk}h_{k\ell}$) that have been assigned to source j. There are K indexes in total that we partition to the J sources with $\mathcal{K} = \{\mathcal{K}_j\}_{j=1}^{J}$. All sources coefficients are assumed to be (statistically) independent.

Local Gaussian Composite Model (LGcM) with NMF An interesting way to arrive at (1.10) is to introduce the *source components* $\{c_{k,f\ell}\}_{k=1}^{K}$, that are also hidden random variables [Févotte 09] and let $c_{k,f\ell}$ follow a complex-Normal PDF:

$$p(c_{k,f\ell}) = \mathcal{N}_c\left(c_{k,f\ell}; 0, u_{k,f\ell}\right), \tag{1.11}$$

[5] Of course, the posterior of a subset depends on all other subsets and so they are updated in alternation.

with factorised variance:

$$u_{k,f\ell} = w_{fk}h_{k\ell}. \tag{1.12}$$

All components are assumed to be independent. Defining $s_{j,f\ell}$ as the sum of the $\mathcal{K}_j$ source components:

$$s_{j,f\ell} = \sum_{k\in\mathcal{K}_j} c_{k,f\ell}, \tag{1.13}$$

results again in (1.10) [Févotte 09].

General LGcM In LGcM, extensive research has been done for the NMF parameters $w_{fk}, h_{k\ell}$. In [Lee 01, Smaragdis 03] they are treated as model parameters in a deterministic model. Alternatively, as for example in [Virtanen 08, Févotte 09, Bertin 10, Hoffman 10, Ozerov 12] the NMF parameters are considered as hidden random variables in a probabilistic model. Also, multiple types of constraints, such as harmonicity and sparsity have been placed on them through prior distributions [Virtanen 07, Ozerov 11].

1.4.3 A State of the Art EM for Audio Source Separation

The LGcM-with-NMF enables for source separation from a single channel mixture.[6] To address MASS the LGcM-with-NMF is combined with a mixing model, as in Section 1.2.2) [Ozerov 10, Arberet 10]. We now present in detail [Ozerov 10] as it is our source of inspiration for the designs of this thesis.

Multichannel Mixtures with LGcM In [Ozerov 10] the source coefficients are hidden random variables modeled with LGcM-with-NMF[7]. Writing (1.13) in vector form:

$$s_{j,f\ell} = \sum_{k\in\mathcal{K}_j} c_{k,f\ell} \Leftrightarrow \mathbf{s}_{f\ell} = \mathbf{G}\mathbf{c}_{f\ell}, \tag{1.14}$$

where the binary matrix $\mathbf{G} \in \mathbb{N}^{J\times K}$ has entries $G_{jk} = 1$ if $k \in \mathcal{K}_j$, and $G_{jk} = 0$ otherwise. The observation model (1.4) is written as:

$$p(\mathbf{x}_{f\ell}|\mathbf{s}_{f\ell}) = \mathcal{N}_c\left(\mathbf{x}_{f\ell}; \mathbf{A}_f\mathbf{s}_{f\ell}, \mathrm{v}_f\mathbf{I}_I\right), \tag{1.15}$$

with $\mathbf{A}_f, \mathrm{v}_f$ model parameters to be estimated.

[6] In practice LGcM-with-NMF separates (the PSD) of $s_{j,f\ell}$ into source components $\{c_{k,f\ell}\}_{k\in\mathcal{K}_j}$.

[7] The number of sources J, components K, and the partition $\mathcal{K}_j$ are known in advance and are fixed.

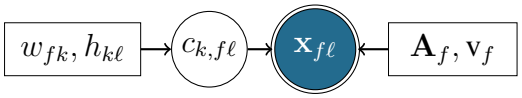

Figure 1.3: Graphical model of the model of [Ozerov 10]. Latent variables are shown as circles, observations as double circles, deterministic parameters with rectangles.

The Generative Model The hidden variables are $\mathcal{H} = \{\mathbf{s}_{f\ell}, \mathbf{c}_{f\ell}\}_{f,\ell=1}^{F,L}$. The model parameters to be estimated are $\theta = \{\mathbf{A}_f, \mathrm{v}_f, u_{k,f\ell}\}_{f,\ell,k=1}^{F,L,K}$. The observations are assumed independent over f, ℓ. Therefore the complete data distribution writes:

$$p\left(\mathcal{H}, \mathbf{x}_{1:F1:L}; \theta\right) = \prod_{f,\ell=1}^{F,L} p(\mathbf{x}_{f\ell}|\mathbf{s}_{f\ell}) \prod_{f,\ell,k=1}^{F,L,K} p(c_{k,f\ell}). \tag{1.16}$$

A *graphical model* for (1.16) can be seen in Fig. 1.3.

E step The posterior distribution of the component vector $p(\mathbf{c}_{f\ell}|\mathbf{x}_{1:F1:L})$ equals the product of all terms of the complete data distribution (1.16) that depend on $\mathbf{c}_{f\ell}$:[8] [9]

$$p(\mathbf{c}_{f\ell}|\mathbf{x}_{1:F1:L}) \propto p(\mathbf{x}_{f\ell}|\mathbf{s}_{f\ell}) \prod_{k=1}^{K} p(c_{k,f\ell}) = \mathcal{N}_c\left(\mathbf{c}_{f\ell}; \hat{\mathbf{c}}_{f\ell}, \mathbf{\Sigma}_{f\ell}^{\eta c}\right), \tag{1.17}$$

with posterior covariance matrix $\mathbf{\Sigma}_{f\ell}^{\eta c}$ and mean vector $\hat{\mathbf{c}}_{f\ell}$ found with:

$$\mathbf{\Sigma}_{f\ell}^{\eta c} = \left[\mathrm{diag}_K\left(\frac{1}{u_{k,f\ell}}\right) + \mathbf{G}^\top \frac{\mathbf{A}_f^{\mathrm{H}}\mathbf{A}_f}{\mathrm{v}_f}\mathbf{G}\right]^{-1}, \tag{1.18}$$

$$\hat{\mathbf{c}}_{f\ell} = \mathbf{\Sigma}_{f\ell}^{\eta c}\mathbf{G}^\top \mathbf{A}_f^{\mathrm{H}} \frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}, \tag{1.19}$$

As shown in the Appendix, the posterior PDF of $\mathbf{s}_{f\ell}$ is also complex-Gaussian:

$$p(\mathbf{s}_{f\ell}|\mathbf{x}_{1:F1:L}) = \mathcal{N}_c\left(\mathbf{s}_{f\ell}; \hat{\mathbf{s}}_{f\ell}, \mathbf{\Sigma}_{f\ell}^{\eta s}\right), \tag{1.20}$$

with covariance matrix $\mathbf{\Sigma}^{\eta s}$ and mean vector $\hat{\mathbf{s}}_{f\ell}$:

$$\mathbf{\Sigma}_{f\ell}^{\eta s} = \left[\mathrm{diag}_J\left(\frac{1}{\sum\limits_{k\in\mathcal{K}_j} u_{k,f\ell}}\right) + \frac{\mathbf{A}_f^{\mathrm{H}}\mathbf{A}_f}{\mathrm{v}_f}\right]^{-1}, \tag{1.21}$$

$$\hat{\mathbf{s}}_{f\ell} = \mathbf{\Sigma}_{f\ell}^{\eta s}\mathbf{A}_f^{\mathrm{H}} \frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}, \tag{1.22}$$

which is a typical Wiener filtering estimator for the sources.

[8] Note that $\prod_{k=1}^{K} p(c_{k,f\ell}) = \mathcal{N}_c\left(\mathbf{c}_{f\ell}; \mathbf{0}_K, \mathrm{diag}_K\left(u_{k,f\ell}\right)\right)$.

[9] Notice also $p(\mathbf{x}_{f\ell}|\mathbf{s}_{f\ell}) = p(\mathbf{x}_{f\ell}|\mathbf{G}\mathbf{c}_{f\ell}) = \mathcal{N}_c\left(\mathbf{x}_{f\ell}; \mathbf{A}_f\mathbf{G}\mathbf{c}_{f\ell}, \mathrm{v}_f\mathbf{I}_I\right)$.

Algorithm 1 [Ozerov 10].

input $\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}$, binary matrix $\mathbf{G}$, initial parameters θ.
repeat

E step

E-$\mathbf{s}_{f\ell}$ *step*: Compute $\mathbf{\Sigma}_{f\ell}^{\eta s}$ with (1.21), then $\hat{\mathbf{s}}_{f\ell}$ with (1.22), and then $\mathbf{Q}_{f\ell}^{\eta s}$ with (1.24).
E-$\mathbf{c}_{f\ell}$ *step*: Compute $\Sigma_{kk,f\ell}^{\eta c}$ with (1.18), $\hat{c}_{k,f\ell}$ with (1.19). Then $Q_{kk,f\ell}^{\eta c}$ with (1.25).

M step

Mixing filter: Update $\mathbf{A}_f$ with (1.26), then v_f with (1.27).
NMF: Update w_{fk} with (1.28), then $h_{k\ell}$ with (1.29).

until convergence
return the estimated source images by applying inverse STFT on $\{A_{ij,f}\hat{s}_{j,f\ell}\}_{f,\ell=1}^{F,L}$.

M step Replacing (1.16) into (1.6) and discarding constants, $\mathcal{L}(\theta)$ writes:

$$\mathcal{L}(\theta) = \sum_{f,\ell=1}^{F,L}\left(-I\log(\mathrm{v}_f) - \frac{1}{\mathrm{v}_f}\mathrm{tr}\Big\{\mathbf{x}_{f\ell}\mathbf{x}_{f\ell}^{\mathrm{H}} - \mathbf{A}_f\hat{\mathbf{s}}_{f\ell}\mathbf{x}_{f\ell}^{\mathrm{H}} - \mathbf{x}_{f\ell}\hat{\mathbf{s}}_{f\ell}^{\mathrm{H}}\mathbf{A}_f^{\mathrm{H}} + \mathbf{A}_f\mathbf{Q}_{f\ell}^{\eta s}\mathbf{A}_f^{\mathrm{H}}\Big\}\right) + \sum_{f,\ell,k=1}^{F,L,K}\left(-\log(w_{fk}h_{k\ell}) - \frac{Q_{kk,f\ell}^{\eta c}}{w_{fk}h_{k\ell}}\right), \tag{1.23}$$

where $u_{k,f\ell}$ is replaced with (1.12). $\mathbf{Q}_{f\ell}^{\eta s} = \mathbb{E}_{q(\mathbf{s}_{f\ell})}\left[\mathbf{s}_{f\ell}\mathbf{s}_{f\ell}^{\mathrm{H}}\right]$, $Q_{kk,f\ell}^{\eta c} = \mathbb{E}_{q(\mathbf{c}_{f\ell})}\left[|c_{k,f\ell}|^2\right]$ are second order moments with respective expressions:

$$\mathbf{Q}_{f\ell}^{\eta s} = \mathbf{\Sigma}_{f\ell}^{\eta s} + \hat{\mathbf{s}}_{f\ell}\hat{\mathbf{s}}_{f\ell}^{\mathrm{H}}, \tag{1.24}$$

$$Q_{kk,f\ell}^{\eta c} = \Sigma_{kk,f\ell}^{\eta c} + |\hat{c}_{k,f\ell}|^2. \tag{1.25}$$

Differentiating (1.23) with respect to $\mathbf{A}_f$ [Hjorungnes 07] and cancelling gives:

$$\mathbf{A}_f = \left(\sum_{\ell=1}^{L}\mathbf{x}_{f\ell}\hat{\mathbf{s}}_{f\ell}^{\mathrm{H}}\right)\left(\sum_{\ell=1}^{L}\mathbf{Q}_{f\ell}^{\eta s}\right)^{-1}. \tag{1.26}$$

Similarly, maximising (1.23) with respect to v_f gives the update rule:

$$\mathrm{v}_f = \frac{1}{LI}\sum_{\ell=1}^{L}\left(\mathbf{x}_{f\ell}^{\mathrm{H}}\mathbf{x}_{f\ell} - 2\mathfrak{Re}\big\{\mathbf{x}_{f\ell}^{\mathrm{H}}\mathbf{A}_f\hat{\mathbf{s}}_{f\ell}\big\} + \mathrm{tr}\big\{\mathbf{Q}_{f\ell}^{\eta s}\mathbf{A}_f^{\mathrm{H}}\mathbf{A}_f\big\}\right). \tag{1.27}$$

Updating v_f is of great importance, and can affect tremendously the quality of the resulting audio signals [Ozerov 10].

Maximising (1.23) with respect to $w_{fk}, h_{k\ell}$ is non-convex. Therefore (1.23) is optimized for w_{fk} keeping $h_{k\ell}$ fixed and vice versa, giving the update rules [Févotte 09]:

$$w_{fk} = \frac{1}{L} \sum_{\ell=1}^{L} \frac{Q^{\eta c}_{kk,f\ell}}{h_{k\ell}}, \tag{1.28}$$

$$h_{k\ell} = \frac{1}{F} \sum_{f=1}^{F} \frac{Q^{\eta c}_{kk,f\ell}}{w_{fk}}. \tag{1.29}$$

The fact that (1.29) considers all frequencies together provides the ability to LGcM-with-NMF of not introducing permutations of the sources (across frequencies). The complete EM algorithm of [Ozerov 10] is given in Algorithm 1.[10]

Estimation of source images from EM The mixing filters and the source signals are always recovered up to a scale factor, which in STFT MASS is frequency dependent. In this thesis, we assess the separation performance of a method using the estimated source images. For example, in [Ozerov 10] after the EM has converged we calculate the estimated j^{th} source image by applying the inverse STFT with overlap-add to $\{A_{ij,f}\hat{s}_{j,f\ell}\}_{f,\ell=1}^{F,L}$.

1.5 CONTRIBUTIONS OF THIS THESIS

In this thesis we investigated the MASS problem of multichannel convolutive audio mixtures. The novelties of this thesis extend in three perspectives that are respectively presented on the three core chapters of this manuscript: In Chapter 2 we propose a more flexible alternative for LGcM-with-NMF where the source prior PSD becomes full rank. In Chapter 3 we study the MASS for time-varying convolutive audio mixtures. In Chapter 4 we design a joint algorithm to simultaneously solve MASS and audio *diarisation* tasks. The three core chapters are followed with a conclusion Chapter 5, where we also discuss the overall material of the manuscript and express various remaining challenges and future directions.

Overall, the three contributions can be viewed as different extensions of [Ozerov 10], as they are presented in this manuscript as three independent models. Nevertheless, they are complementary and as such any composition of them is straightforward and is envisioned for future research.

Source Modeling In Chapter 2, we inspire from LGcM-with-NMF and propose a new statistical model for the power spectral density (PSD) of an audio signal and apply it to MASS. To this aim, we derive a vEM algorithm for parameter estimation and source inference. We model the source signals with the LGcM and we propose to model the variance $u_{k,f\ell}$ of each source component with an inverse-Gamma distribution, whose scale

[10] The index of iterations of the EM has been dropped.

parameter is factorised as in a rank-1 NMF. We name this model *Nonnegative Matrix Factorization through inverse Gamma* (NMFiG). NMFiG advances the theory of LGcM-with-NMF by modelling the audio signal with the same (number of) parameters as the NMF but without actually factorising the audio signal's spectrogram. NMFiG also includes a *relevance determination mechanism* to weigh the importance of the individual LGcM components. We benchmark the proposed vEM with the state of the art. Our results have been published in [Kounades-Bastian 16a].

Source Separation of Moving Sound Sources In Chapter 3 we explore MASS for time-varying audio mixtures, which arise when the mixing filters are time-varying. Time-varying mixing filters can describe moving sources, moving microphones, or other changes in the recording environment such as opening of a window, or a blind, etc. Addressing time-varying mixtures is an important feature for a real-world MASS method. To this aim, we allow the mixing matrix in (1.4) to vary with the time frame. To keep the parameter space compact we introduce a Markov chain linking the mixing matrices of succesive time frames. The sources are modeled with LGcM-with-NMF. We derive a vEM algorithm that uses a Kalman smoother to infer the time-varying mixing matrix and the source signals. Extensive experiments on simulated and real recordings show that the proposed method outperforms the block-wise adaptation of two state of the art MASS methods for time-invariant mixtures. Our results have been published in [Kounades-Bastian 15, Kounades-Bastian 16b].

Joint Audio Diarisation and Audio Source Separation Audio diarisation is the labeling of the audio mixture with the sources (for example the speakers) that are emitting at each time [Anguera Miro 12]. Audio diarisation is closely related with MASS, and in Chapter 4 we propose a joint formulation of these two problems. We propose a generative model to perform jointly MASS and audio diarisation of convolutive audio mixtures by augmenting (1.4) with a activity labeling mechanism for every source at the STFT frame level. We model the sources with the LGcM-with-NMF and derive an EM algorithm to infer the label (diarisation) and the separated source signals. The diarisation is aided by a Hidden Markov Model (HMM). The proposed EM shows separation performance comparable with [Ozerov 10], while outperforming a state of the art speaker diarisation pipeline. Our results have been published in [Kounades-Bastian 17].

CHAPTER 2

A GENERATIVE MODEL FOR SPECTROGRAM FACTORISATION

We inspire from the LGcM-with-NMF design, and propose a statistical model for the PSD of an audio signal. The heart of this model is a novel setting of the variance of the LGcM components. We assume the variance of a LGcM component to be a latent random variable, following an inverse-Gamma distribution, whose *scale parameter* is factorised as a rank-1 model. This way we inherit all useful properties of the LGcM-with-NMF but without restricting the source PSD matrix to be of low-rank. We name this new model *Nonnegative Matrix Factorization with inverse-Gamma* (NMFiG). We include the proposed formulation to a MASS framework. We derive a vEM algorithm for estimation of the model parameters and source inference. We evaluate its performance on separating real-world and simulated underdetermined mixtures of speech. NMFiG shows a benefit in source separation performance compared to a state of the art LGcM-with-NMF technique. Finally we draw our insights on the ability of NMFiG to weigh the importance of the LGcM components.

2.1 INTRODUCTION

In LGcM-with-NMF the variance of a component is considered to factorise over frequency and time, as in (1.12). In the present chapter we propose an extension of LGcM-with-NMF where the component variance becomes a latent random variable and no factorisation is applied on her. We envision the extension to be an alternative for LGcM-with-NMF in MASS of speech mixtures. As our interest is in multichannel mixtures, we include the proposed LGcM variant in the MASS framework of [Ozerov 10]. We derive the associated vEM, because the E step of an exact EM is not analytically tractable.

The main feature of the proposed LGcM variant is the design of the prior distribution placed on the component variance $u_{k,f\ell}$. We choose this prior from the family of Inverse

Gamma (IG) distributions; a family that is common in Bayesian NMF [Cemgil 09]. The IG distribution is defined by two non-negative parameters, a shape and a scale. We choose to parametrise the scale with a factorised rank-1 model, reminiscent of NMF, and let the shape control the participation of the specific component. Hence there is a single shape parameter per component that does not depend neither on frequency nor on the frame. Recall that the variance of a source in LGcM is the sum of the variances of its components. This way we make the proposed parametrisation to have almost the same number of model parameters as the LGcM-with-NMF (up to few additional shape parameters).

We now detail NMFiG and include it in the multichannel framework of [Ozerov 10]. Then we derive the associated vEM, that we name *variational EM with NMFiG* (vEMiG). In Section 2.4, we benchmark the vEMiG against [Ozerov 10] on MASS tasks of simulated and real-world underdetrmined mixtures of speech.

2.2 MULTICHANNEL NONNEGATIVE MATRIX FACTORIZATION WITH INVERSE GAMMA

We work under the narrow-band assumption, which allows us to write a time-invariant convolutive mixture of I channels in the STFT with (1.4). Then, to express the mixture probabilistically, we use (1.15) as in [Ozerov 10].

2.2.1 THE LGCM SOURCE MODEL WITH INVERSE GAMMA

We consider the source coefficients as hidden variables following LGcM, with (1.14). In LGcM-with-NMF the variance $u_{k,f\ell}$ is assumed to factorise over f and ℓ (see (1.12)). That factorisation may introduce artefacts if applied to intricate audio signals, such as speech. We propose here to relax this assumption, by letting $u_{k,f\ell}$ to be a hidden random variable. Therefore (1.11) naturally becomes:

$$p(c_{k,f\ell}|u_{k,f\ell}) = \mathcal{N}_c\left(c_{k,f\ell}; 0, u_{k,f\ell}\right). \tag{2.1}$$

We set $u_{k,f\ell}$ to follow an inverse Gamma (IG) distribution:

$$p(u_{k,f\ell}) = \mathcal{IG}\left(u_{k,f\ell}; \gamma_k, \delta_{k,f\ell}\right), \tag{2.2}$$

with the scale-parameter $\delta_{k,f\ell} \in \mathbb{R}_+$ factorised as:

$$\delta_{k,f\ell} = w_{fk} h_{k\ell}, \tag{2.3}$$

with $\gamma_k, w_{fk}, h_{k\ell}$ being non-negative parameters to be estimated.

The key point of (2.3) is to keep the number of model parameters low. Since, having $\delta_{k,f\ell}$ factorised as a rank-1, make the number of parameters be equal (plus few additional γ_k) with those of LGcM-with-NMF.

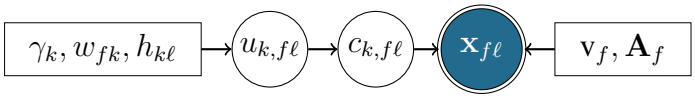

Figure 2.1: Graphical model for MASS with NMFiG. Latent variables are shown as circles, observations as double circles, deterministic parameters with rectangles.

The additional γ_k shape parameters play an important role as they are responsible to suppress irrelevant components, thus balancing the capacity of LGcM. Recall that the assignment of components to sources is known beforehand.

Since there is no factorisation on the variance $u_{k,f\ell}$ of the NMFiG component. The NMFiG component may be able to represent more intricate spectrograms, compared to its analog: the LGcM-with-NMF component (recall (1.12)).

2.2.2 The Complete Data Probability Distribution

Our set of hidden variables $\mathcal{H} = \{\mathbf{s}_{f\ell}, \mathbf{c}_{f\ell}, u_{k,f\ell}\}_{f,\ell,k=1}^{F,L,K}$ consists of the sources, the components, and their PSD. Let all components, and all PSDs to be mutually and individually independent a priori. The complete data PDF writes:

$$p\left(\mathcal{H}, \mathbf{x}_{1:F1:L}; \theta\right) = \prod_{f,\ell=1}^{F,L} p\left(\mathbf{x}_{f\ell}|\mathbf{s}_{f\ell}\right) \prod_{f,\ell,k=1}^{F,L,K} p(c_{k,f\ell}|u_{k,f\ell}) \prod_{f,\ell,k=1}^{F,L,K} p(u_{k,f\ell}). \tag{2.4}$$

The set of model parameters $\theta = \{\mathbf{A}_f, \mathrm{v}_f, w_{fk}, h_{k\ell}, \gamma_k\}_{f,\ell,k=1}^{F,L,K}$ consists of the mixing matrices, the residual noise variance, and the IG parameters. The *graphical model* where we see the prior dependencies of the hidden variables is depicted in Fig. 2.1.

2.3 The vEMiG Algorithm

We develop a variational approximation of the true posterior (see Section 1.4.1):

$$q(\mathcal{H}) = \prod_{f,\ell=1}^{F,L} q(\mathbf{c}_{f\ell}) \prod_{f,\ell,k=1}^{F,L,K} q(u_{k,f\ell}). \tag{2.5}$$

vEMiG consists of an E step and an M step. In the E step we first compute $q(u_{k,f\ell})$ with (1.8), and then compute $q(\mathbf{c}_{f\ell})$ also with (1.8). Interestingly both factors are identified in closed form. In the M step we optimize $\mathcal{L}(\theta)$ given by (1.9), to update θ.

2.3.1 E step

For ease of presentation we partition the E-step in three steps: The E-$u_{k,f\ell}$ step that computes $q(u_{k,f\ell})$, the E-$\mathbf{c}_{f\ell}$ step that computes $q(\mathbf{c}_{f\ell})$, and the E-$\mathbf{s}_{f\ell}$ step that computes $q(\mathbf{s}_{f\ell})$

whose mean vector $\hat{\mathbf{s}}_{f\ell} \in \mathbb{C}^J$ is the MAP estimator for $\mathbf{s}_{f\ell}$.

E-$u_{k,f\ell}$ step Eq. (1.8), replacing (2.4) and discarding constants, writes:[1]

$$q(u_{k,f\ell}) \propto p(u_{k,f\ell}) \exp\Big(\mathbb{E}_{q(\mathbf{c}_{f\ell})}\big[\log p(c_{k,f\ell}|u_{k,f\ell})\big]\Big) \propto \tag{2.6}$$

$$\mathcal{IG}\big(u_{k,f\ell}; g_k, d_{k,f\ell}\big), \tag{2.7}$$

with g_k and $d_{k,f\ell}$ computed with:

$$g_k = \gamma_k + 1, \tag{2.8}$$

$$d_{k,f\ell} = \delta_{k,f\ell} + Q^{\eta c}_{kk,f\ell}. \tag{2.9}$$

Note, that the expectation in (2.6) is:

$$\exp\big(\mathbb{E}_{q(\mathbf{c}_{f\ell})}[\log p(c_{k,f\ell}|u_{k,f\ell})]\big) \propto (u_{k,f\ell})^{-1} \exp\left(-\frac{Q^{\eta c}_{kk,f\ell}}{u_{k,f\ell}}\right), \tag{2.10}$$

with $Q^{\eta c}_{kk,f\ell} = \mathbb{E}_{q(\mathbf{c}_{f\ell})}\left[|c_{k,f\ell}|^2\right] \in \mathbb{R}_+$ provided from E-$\mathbf{c}_{f\ell}$ step. The calculation of $Q^{\eta c}_{kk,f\ell}$ will resolve, after we identify the $q(\mathbf{c}_{f\ell})$ in the next paragraph. Note, that we made no assumption on the functional form of the distribution $q(\mathbf{c}_{f\ell})$.

E-$\mathbf{c}_{f\ell}$ step We use (1.8), but now we are interested in $q(\mathbf{c}_{f\ell})$; replacing (2.4) into the former and discarding any terms not depending on $\mathbf{c}_{f\ell}$, (1.8) writes:

$$q(\mathbf{c}_{f\ell}) \propto p(\mathbf{x}_{f\ell}|\mathbf{s}_{f\ell}) \prod_{k=1}^{K} \exp\big(\mathbb{E}_{q(u_{k,f\ell})}\big[\log p(c_{k,f\ell}|u_{k,f\ell})\big]\big) = \tag{2.11}$$

$$\mathcal{N}_c\left(\mathbf{c}_{f\ell}; \hat{\mathbf{c}}_{f\ell}, \mathbf{\Sigma}^{\eta c}_{f\ell}\right), \tag{2.12}$$

where, using (2.7) and (2.1), we easily find:

$$\exp\big(\mathbb{E}_{q(u_{k,f\ell})}[\log p(c_{k,f\ell}|u_{k,f\ell})]\big) \propto \mathcal{N}_c\left(c_{k,f\ell}; 0, \hat{u}_{k,f\ell}\right), \tag{2.13}$$

with $\hat{u}_{k,f\ell} \in \mathbb{R}_+$ defined:

$$\hat{u}_{k,f\ell} = \left(\mathbb{E}_{q(u_{k,f\ell})}\left[\frac{1}{u_{k,f\ell}}\right]\right)^{-1} = \frac{d_{k,f\ell}}{g_k}. \tag{2.14}$$

The posterior mean vector $\hat{\mathbf{c}}_{f\ell}$ and covariance matrix $\mathbf{\Sigma}^{\eta c}_{f\ell}$ are obtained with:

$$\mathbf{\Sigma}^{\eta c}_{f\ell} = \left[\mathrm{diag}_K\left(\frac{1}{\hat{u}_{k,f\ell}}\right) + \mathbf{G}^\top \frac{\mathbf{A}_f^{\mathrm{H}}\mathbf{A}_f}{\mathrm{v}_f}\mathbf{G}\right]^{-1}, \tag{2.15}$$

$$\hat{\mathbf{c}}_{f\ell} = \mathbf{\Sigma}^{\eta c}_{f\ell}\mathbf{G}^\top \mathbf{A}_f^{\mathrm{H}} \frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}, \tag{2.16}$$

Recall here, $Q^{\eta c}_{kk,f\ell}$ needed for (2.10). Easily, $Q^{\eta c}_{kk,f\ell}$ is computed with (1.25) although using $\Sigma^{\eta c}_{kk,f\ell}$ from (2.15) and $\hat{c}_{k,f\ell}$ from (2.16).

[1] We drop from (1.8) any multiplicative factor that does not depend on the variable at stake, here $u_{k,f\ell}$.

E-$\mathbf{s}_{f\ell}$ step Due to LGcM, and as shown in the appendix, $q(\mathbf{s}_{f\ell})$is:

$$q(\mathbf{s}_{f\ell}) = \mathcal{N}_c\left(\mathbf{s}_{f\ell}; \hat{\mathbf{s}}_{f\ell}, \mathbf{\Sigma}_{f\ell}^{\eta s}\right), \tag{2.17}$$

with covariance matrix $\mathbf{\Sigma}_{f\ell}^{\eta s}$ and mean vector $\hat{\mathbf{s}}_{f\ell}$ given:

$$\mathbf{\Sigma}_{f\ell}^{\eta s} = \left[\text{diag}_J\left(\frac{1}{\sum\limits_{k\in\mathcal{K}_j}\hat{u}_{k,f\ell}}\right) + \frac{\mathbf{A}_f^{\text{H}}\mathbf{A}_f}{\mathrm{v}_f}\right]^{-1}, \tag{2.18}$$

$$\hat{\mathbf{s}}_{f\ell} = \mathbf{\Sigma}_{f\ell}^{\eta s}\mathbf{A}_f^{\text{H}}\frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}. \tag{2.19}$$

of course, $\hat{\mathbf{s}}_{f\ell}$ is the MAP estimator for the separated sources provided by vEMiG. Juxtaposing (2.18) and (1.21), both have similar forms, but in the latter the component PSD $u_{k,f\ell}$ is a rank-1 parameter, where in the former is $\hat{u}_{k,f\ell}$ (an unfactorised expectation). This major difference adds flexibility on the Wiener filters that provide $\hat{\mathbf{s}}_{f\ell}$.

2.3.2 M STEP

In this section we develop the updates for the parameters in θ.

M-$\mathbf{A}_f, \mathbf{v}_f$ step Because our mixing equation is (1.15), that is the same with [Ozerov 10], we obtain the updates: (1.26) for $\mathbf{A}_f$ and (1.27) v_f, as derived in Section 1.4.3. The second order moment of the sources $\mathbf{Q}_{f\ell}^{\eta s}$ is also given with (1.24), but with $\mathbf{\Sigma}_{f\ell}^{\eta s}$ from (2.18) and $\hat{\mathbf{s}}_{f\ell}$ from (2.19).

M-IG step The ECDLL for the IG parameters is given by (1.9), by replacing (2.2):

$$\mathcal{L}\left(\{w_{fk}, h_{k\ell}, \gamma_k\}_{f,\ell,k=1}^{F,L,K}\right) = \sum_{f,\ell,k=1}^{F,L,K}\mathbb{E}_{q(u_{k,f\ell})}\left[\log\mathcal{IG}\left(u_{k,f\ell}; \gamma_k, \delta_{k,f\ell}\right)\right] = \tag{2.20}$$

$$\sum_{f,\ell,k=1}^{F,L,K}\left(\gamma_k\log\left(w_{fk}h_{k\ell}\right) - \log\Gamma(\gamma_k) - \gamma_k\Big(\log(d_{k,f\ell}) - \psi(g_k)\Big) - \frac{w_{fk}h_{k\ell}}{\hat{u}_{k,f\ell}}\right), \tag{2.21}$$

with $\psi()$ the digamma function. Maximising (2.21) for w_{fk} (fixing other terms) results:

$$w_{fk} = \frac{L\gamma_k}{\sum\limits_{\ell=1}^{L}\frac{h_{k\ell}}{\hat{u}_{k,f\ell}}}. \tag{2.22}$$

Maximising (2.21) now for $h_{k\ell}$, gives a similar update:

$$h_{k\ell} = \frac{F\gamma_k}{\sum\limits_{f=1}^{F}\frac{w_{fk}}{\hat{u}_{k,f\ell}}}. \tag{2.23}$$

Interestingly, the sum now appears on the denominator, whereas in the standard LGcM-with-NMF appears in the numerator (see (1.28) and (1.29)).

Differentiating (2.21) for γ_k and setting the result to zero, results in the equation:

$$\sum_{f,\ell=1}^{F,L}\left(\log\left(\frac{w_{fk}h_{k\ell}}{d_{k,f\ell}}\right)-\psi(\gamma_k)+\psi\left(g_k\right)\right)=0. \tag{2.24}$$

Eq. (2.24) is non-linear on γ_k. To solve (2.24) we replace $g_k, d_{k,f\ell}$ with their respective expressions (2.8), (2.9). Then (2.24) writes:

$$\sum_{f,\ell=1}^{F,L}\left(-\log\left(1+\frac{Q^{\eta c}_{kk,f\ell}}{w_{fk}h_{k\ell}}\right)-\psi(\gamma_k)+\psi(\gamma_k+1)\right)=0. \tag{2.25}$$

Using the reflection formula: $\psi(\gamma_k+1)=\psi(\gamma_k)+\frac{1}{\gamma_k}$ [Abramowitz 65], (2.25) writes:

$$\sum_{f,\ell=1}^{F,L}\left(-\log\left(1+\frac{Q^{\eta c}_{kk,f\ell}}{w_{fk}h_{k\ell}}\right)-\cancel{\psi(\gamma_k)}+\cancel{\psi(\gamma_k)}+\frac{1}{\gamma_k}\right)=0. \tag{2.26}$$

Solving (2.26) for γ_k is now closed form:

$$\gamma_k=\frac{FL}{\sum\limits_{f,\ell=1}^{F,L}\log\left(1+\frac{Q^{\eta c}_{kk,f\ell}}{w_{fk}h_{k\ell}}\right)}. \tag{2.27}$$

Eq. (2.27) is an ML estimator for the shape parameter of an IG probability distribution.

2.3.3 Implementing vEMiG

In Algorithm 2 we give the vEMiG algorithm as it is implemented. The order of execution of the respective E and M steps is chosen empirically.

2.4 Experimental Study

In this section we benchmark vEMiG on MASS tasks of underdetermined convolutive stereo mixtures of speech. In specific, we evaluate our method in separating $J=3$ speech signals from artificially-generated convolutive stero $I=2$ mixtures, and we present average results over 8 realizations with different source signals. As baseline method we choose [Ozerov 10], as it is the closest in spirit to our method. Initially, we describe the simulation setup and the mixture configuration. Then we explain how we choose initial values for the parameters θ for the vEM and for the baseline. We evaluate the separated signals by the two methods, quantitatively using standard MASS measures [Vincent 06]. We end with a subsection with insights on by-properties of the NMFiG model.

Algorithm 2. vEMiG: A vEM for multichannel source separation with NMFiG.

input $\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}$, binary matrix $\mathbf{G}$, initial parameters θ.
initialise: The IG parameters $\left\{\hat{u}_{k,f\ell}\right\}_{f,\ell,k=1}^{F,L,K}$.
repeat

E step

E-$\mathbf{s}_{f\ell}$ *step*: Compute $\mathbf{\Sigma}_{f\ell}^{\eta s}$ with (2.18) and $\hat{\mathbf{s}}_{f\ell}$ with (2.19). Then $\mathbf{Q}_{f\ell}^{\eta s}$ with (1.24).
E-$\mathbf{c}_{f\ell}$ *step*: Compute $\Sigma_{kk,f\ell}^{\eta c}$ with (2.15), $\hat{c}_{k,f\ell}$ with (2.16), then $Q_{kk,f\ell}^{\eta c}$ with (1.25).
E-$u_{k,f\ell}$ *step*: Compute g_k with (2.8), $d_{k,f\ell}$ with (2.9), then $\hat{u}_{k,f\ell}$ with (2.14).

M step

M-IG step: Update w_{fk} with (2.22), then $h_{k\ell}$ with (2.23), and then γ_k with (2.27).
M-$\mathbf{A}_f$ *step*: Update $\mathbf{A}_f$ with (1.26).
M-v_f *step*: Update v_f with (1.27).

until convergence
return the estimated source images by applying inverse STFT on $\{A_{ij,f}\hat{s}_{j,f\ell}\}_{f,\ell=1}^{F,L}$.

2.4.1 INITIALIZING THE MODEL PARAMETERS

LGcM models have a large number of parameters to be estimated. Both vEMiG and the baseline method are iterative optimization techniques. As such, they can *stuck in local optima* of the ECDLL, if their parameters are initialized improperly. For LGcM-witn-NMF based MASS methods it has been observed [Ozerov 10, Arberet 10] that the initial values for the NMF parameters $\{w_{fk}, h_{k\ell}\}_{f,\ell=1}^{F,L}$ are of paramount importance for an acceptable quality of source separation to be achieved. In this thesis we use two initialization strategies which we describe now in detail, and refer here in subsequent chapters.

Semi-blind initialization of NMF parameters The NMF parameters $w_{fk}, h_{k\ell}$ of a given source j are initialized by applying the KL-NMF algorithm [Févotte 09], with $K_j = 20$, to the power spectrogram of a corrupted version of source j. The corrupted version is made by adding to source signal j, scaled versions of all other interfering sources. The corruption is controlled by a signal-to-noise ratio (SNR) R. We test three different levels of corruption, namely $R = 20$dB, $R = 10$dB and 0dB. With 0 dB meaning here equal power of the desired signal ($s_j(t)$), and the sum of all interfering source signals. Clearly $R = 20$ dB is a quite favorable initialization aiming to show the upper bound of EM's performance, whereas $R = 0$ dB approaches the realism. This NMF initialization process is applied independently to all sources $j \in [1, J]$. The calculated NMF initial parameters are used for both the vEM and the baseline method.

Blind initialization of NMF parameters For a blind initialization procedure of the NMF parameters, we use a state of the art blind source separation method to provide (initial) estimates for the source spectrograms. NMF decomposition is then applied on those spectrograms to obtain the initial NMF parameters. As state of the art blind source separation method we chose the sound source localization method of [Dorfan 15], which is a good representative of recently proposed probabilistic methods based on mixture models of acoustic feature distribution parametrised by source location, see for example [Mandel 10, May 11, Woodruff 12, Traa 14]. The method of [Dorfan 15] relies on a *mixture of complex Gaussian distributions* (CGMM) that is used to compare the measured *normalized relative transfer function* (NRTF) at a pair of microphones with the expected NRTF as predicted by a source at a candidate position[2]. After identifying the parameters of the CGMM with an EM algorithm. Selecting the J first maxima of the prior probabilities amounts to localize the J sources. Selecting the TF points that have been clustered at each of those J maxima (after comparing the posterior probabilities of the CGMM), provides binary masks for the J sources. Then by applying those masks onto the mixture STFT we obtain the source image STFT coefficients for every source. Then we take the absolute squared values of the estimated source image of source j, average them across channels, supply them (as $F \times L$ matrix) to the KL-NMF algorithm (with $K_j = 20$) [Févotte 09]. Thus obtaining initial NMF parameters. The initial NMF parameters are provided to both the vEM and the baseline method.

Other parameters: For both the vEMiG and the baseline we set $\mathbf{A}_f = \mathbf{1}, \forall f$ (an $I \times J$ matrix filled with ones), and set $\mathrm{v}_f = \frac{10^3}{FLI} \sum_{f,\ell=1}^{F,L} \mathbf{x}_{f\ell}^{\mathrm{H}} \mathbf{x}_{f\ell}$, $\forall f$. We run 20 iterations. For the additional parameters of vEMiG we initialise: $\gamma_k = 1$ and $\hat{u}_{k,f\ell} = w_{fk} h_{k\ell}$, $\forall f, \ell, k$.

2.4.2 SIMULATION SETUP

The convolutive mixtures were generated using a database of *binaural impulse responses* (BRIR) [Hummersone 13] as mixing filters, and (single channel) speech signals as the source signals (they were 2s signals sampled at 16kHz), randomly chosen from the TIMIT database [Garofolo 93]. The BRIRs were recorded with a dummy head equipped with $I = 2$ microphones (one per each ear), placed in a large theater-like room of dimensions $23.5\,\mathrm{m} \times 18.8\,\mathrm{m} \times 4.6\,\mathrm{m}$ with reverberation time $\mathrm{RT}_{60} \approx 0.68\,\mathrm{s}$ [Hummersone 13]. The original BRIRs had $16,000$ taps each, but we truncated them keeping only the leading 512 taps because of memory limitations[3]. The BRIRs were sampled at azimuthal points from -90° to 90° with spacing of 5°, on a circle of radius 1.5m and center the dummy head. We selected BRIRs for $J = 3$ distinct azimuths, namely for $-85^\circ, -20^\circ, 60^\circ$. We convolved each of the single channel source-signals with the pair of BRIRs for that respective azimuth. In this way we displayed a source signal at a spatial position. The source-images were then summed together to provide the mix signal. We then calculated

[2]There is one CGMM component for each candidate source position on a predefined grid. The grid is defined in advance based on a direct-path propagation model.

[3]Hence, the effective RT_{60} is somewhat reduced.

Table 2.1: Quantitative Audio Source Separation Evaluation of NMFiG.

	R	20 dB			10 dB			0 dB			Blind Init.		
Metric	Method	s_1	s_2	s_3	s_1	s_2	s_3	s_1	s_2	s_3	s_1	s_2	s_3
SDR	NMFiG	**11.0**	**9.6**	**8.7**	**9.7**	**8.1**	**8.0**	**5.2**	**4.9**	**3.8**	**5.8**	6.4	**2.8**
	[Ozerov 10]	10.1	8.5	8.2	9.5	7.7	7.5	4.7	3.0	3.5	5.1	**6.7**	2.5
	[Dorfan 15]	-	-	-	-	-	-	-	-	-	4.3	4.1	1.7
SIR	NMFiG	**15.8**	**15.9**	**14.2**	**13.7**	**13.3**	**12.7**	**7.2**	**7.9**	**5.2**	7.3	12.5	4.1
	[Ozerov 10]	14.8	14.8	12.9	13.1	12.8	11.7	6.6	7.1	4.6	**7.8**	**12.6**	4.1
	[Dorfan 15]	-	-	-	-	-	-	-	-	-	1.3	7.4	**8.9**
SAR	NMFiG	**15.2**	15.5	12.6	**15.6**	16.2	**12.9**	**11.1**	**12.0**	**10.7**	**13.3**	11.9	9.6
	[Ozerov 10]	14.4	**16.5**	**12.7**	15.5	**16.4**	12.7	10.6	11.7	9.6	13.2	**12.7**	**9.8**
	[Dorfan 15]	-	-	-	-	-	-	-	-	-	12.0	7.0	8.4

the STFT of the mixture, using a 512-taps sine-wave analysis window with 50% overlap (of samples) between frames and provided to the algorithms.

2.4.3 Results on Audio Source Separation

Table 2.2: Input scores for the mixture.

	SDR			SIR			SAR		
	s_1	s_2	s_3	s_1	s_2	s_3	s_1	s_2	s_3
Mixture	-0.8	-5.9	-4.6	-0.3	-5.1	-3.7	$+\infty$	$+\infty$	$+\infty$

We evaluated the source separation performance of the vEM against, [Ozerov 10], and against [Dorfan 15] (used alone, i.e. using its binary masking estimates). For performance evaluation, we used standard objective measures for MASS [Vincent 06], that are calculated by comparing the estimated and ground truth source images. The measures are: The *signal-to-distortion* (SDR), *the signal-to-interference* (SIR), and *the signal-to-artefact ratios* (SAR), all in dB. All performance scores are reported in Table 2.1. Every reported value is an average result over 8 mixture realizations. The same azimuthal positions were used for all 8 mixtures, but the speech contents of each were randomly chosen from the TIMIT database. For comparison we report in Table 2.2 the *input scores*[4]

Results with controlled initialization From Table 2.1 we see that for $R = 20$dB the vEM improves the SDR of all sources by at least 11.8dB (least for s_1). The SDR of s_2 increases by 15.5dB (from -5.9dB at the input to 9.6dB after running the vEM). Similar improvements are achieved also from [Ozerov 10], again for s_2 and $R = 20$dB the SDR rises by 14.4dB (from -5.9dB at the input to 8.5dB). The proposed vEM scores higher for all $J = 3$ sources, outperforming [Ozerov 10] by 0.9dB for s_1, 1.1dB for s_2 and 0.5dB for s_3. For $R = 0$dB, we see that all scores are lower, clearly due to the corruption of initialization. Still, a consistent benefit is observed in favor of the vEM, for example

[4]Separation scores calculated using the mixture signal as estimator for every source.

the vEM attains an SDR of 4.9dB for s_2 with the baseline scoring at 3.0dB. The vEM rises the SIR of s_2 by 21dB (from -5.9dB to 15.9dB at $R = 20$dB). In terms of SAR, we observe that it starts from perfect in the input ($+\infty$ as all sources are intact in the mix containing no artefacts) and degrades, as any separating technique introduces some artefacts. In SAR, the scores of vEM and [Ozerov 10] are similar, which possibly happens due to them sharing the same mixing model.

Results with blind initialization In terms of SIR and SAR the results are mitigated. Therefore we focus on SDR as it summarizes the overall quality of the separated signals. We see that the initialization method [Dorfan 15] attains SDRs of 4.3dB for s_1, 4.1dB for s_2, 1.7dB for s_3. Execution of either the vEM or the baseline [Ozerov 10], increases the scores provided by [Dorfan 15]. After initializing the NMF with [Dorfan 15], the vEM improves the SDR scores by: 1.5dB for s_1, 2.3dB for s_2 and 1.1dB for s_3. For s_1, s_3 the vEM outperforms the SDRs of [Ozerov 10]. The overall improvement of SDR is attributed to the use of the NMFiG model in the place of the standard LGcM-with-NMF. And this inspires us to further investigate the potential of NMFiG full-rank PSD modeling, for source separation and beyond.

2.4.4 THE SHAPE HYPERPARAMETER OF INVERSE GAMMA

We asserted earlier that γ_k controls the contribution, of the k-th component in the PSD of j_k-th source. Eq. (2.14) shows that a high (respectively low) value of γ_k decreases (respectively increases) the value of $\hat{u}_{k,f\ell}$. Then $\hat{u}_{k,f\ell}$ contributes in the posterior estimate of $\hat{s}_{j,f\ell}$ via (2.18). As γ_k is shared across f, ℓ, it controls all $\hat{u}_{k,1:F1:L}$ simultaneously. Fig. 2.2 demonstrates experimentally this fact: For $R = 20$dB where components are learned from the true source spectra the vEMIG is able to tell which k's are "relevant" for (2.18) as quantified by extremely small (relevant) or extremely high (irrelevant) estimated values for γ_k. When for $R = 0$dB where the learned components are more corrupted, the vEM is less decisive and yields less extreme values for γ_k. Recall that in both cases γ_k is initialised to 1.

2.5 CONCLUSION

In this chapter we introduced the NMFiG; a new method to model sound source PSD inspired by the LGcM-with-NMF. While in conventional Bayesian NMF, the source PSD is modeled with a NMF for which a prior probability distribution is set, in NMFiG we first model the component PSD with a prior distribution (for instance IG), to later on impose an NMF structure on the scale parameter(s) of the IG prior. We incorporated NMFiG into a MASS fremework, and we derived the associated vEM to infer the source signals. We assessed the performance of the model and the proposed vEM in the challenging task of separating the sound sources from undetermined time-invariant convolutive mixtures of speech signals. The experiments show the interest of the NMFiG when compared to

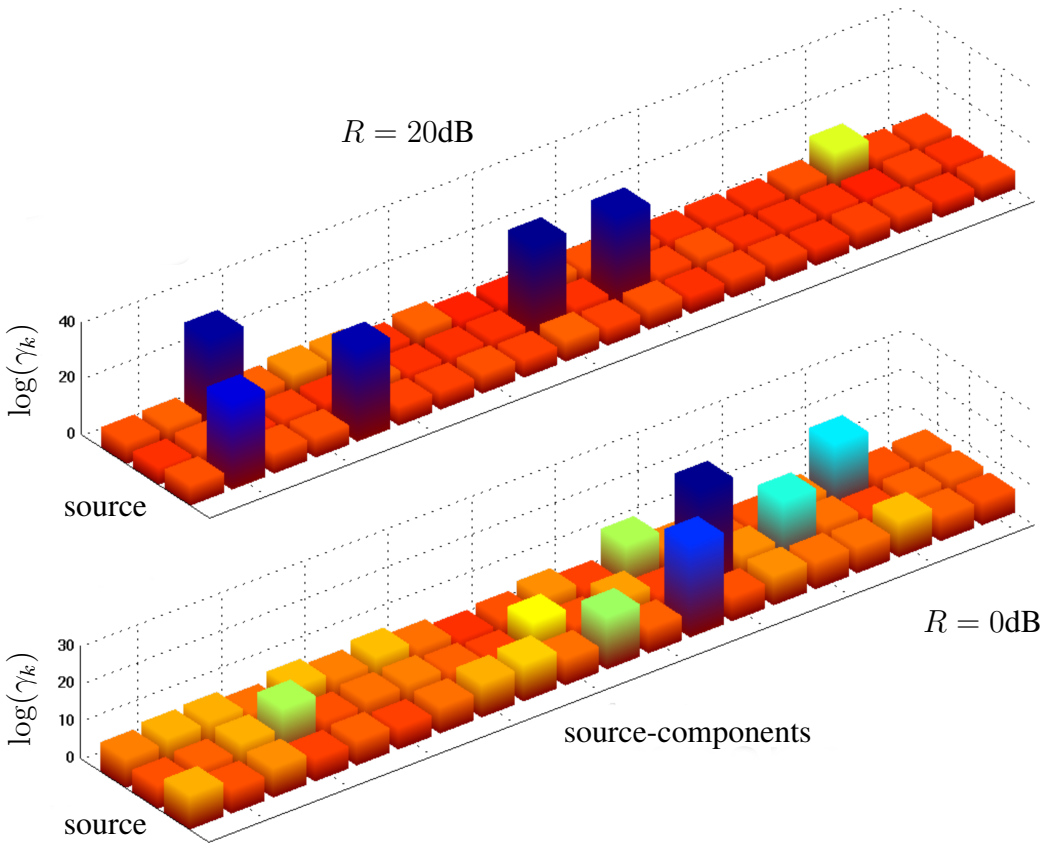

Figure 2.2: Estimated values of $\log(\gamma_k)$, at the last iteration of the vEM applied on the mixtures of Section 2.4.2, with controlled initialization. **Top** $R = 20$dB. **Bottom** $R = 0$dB. A higher value of γ_k decreases the contribution of the corresponding component.

[Ozerov 10]. A qualitative visualization of the estimated values of the shape parameter of the IG, reveals the potential of the NMFiG model to automatically determine the relevant components [Tan 13] of the NMF decomposition. One may envision extending the NMFiG in terms of the excitation-filter NMF model [Ozerov 12].

CHAPTER 3

SOURCE SEPARATION OF TIME-VARYING AUDIO MIXTURES

The chapter addresses the problem of MASS from time-varying convolutive mixtures. Such mixtures can describe movements of the sources and of the sensor-set, and also changes of the environment that happen during the recording, for example opening of a window. We propose a probabilistic framework, based on LGcM-with-NMF, on which we consider the mixing filters to be time-varying, modeled as continuous temporal stochastic processes. We design a vEM algorithm for source separation that uses a Kalman smoother to track and infer the time-varying mixing matrices. Extensive experiments on simulated time-varying convolutive mixtures and and real-world mixtures, of speech, show that the proposed method outperforms a block-wise adaptation of a state-of-the-art time-invariant MASS baseline method.

3.1 INTRODUCTION

In many Human-robot interaction scenarios, there is a strong need to consider mixed speech signals emitted by *moving* speakers, and/or recorded by a *moving* robot, and perturbed by reverberations. More generally, changes in the environment such as door/window opening/closing or curtain pulling must also be accounted for.

All those facts can be represented by the variation over time of the acoustic channel between microphones and sources. The vast majority of works in MASS from convolutive mixtures deal with time-invariant mixing filters. Time-invariant mixing filters are valid when the acoustic channel is time-invariant, which happens when the position of sources and microphones is fixed. In this chapter we consider the mixing filters as time-varying and investigate MASS on such mixtures through a probabilistic formulation.

We start by reviewing the MASS literature for time-varying mixtures and positioning ourselves in. Then, in Section 3.3, we present the proposed probabilistic model. In

Section 3.4 we derive the associated vEM. In Section 3.5 we report the results of the experimental study. In Section 3.6 we conclude, discuss, and give promising future directions.

3.2 Literature Review on Moving Sound Source Separation

Early attempts addressing the separation of time-varying mixtures, consisted in block-wise adaptations of time-invariant methods: The observations (STFT mixture coefficients) are split in blocks of (STFT) frames, and a time-invariant MASS method is applied to each block. Hence, block-wise adaptations assume time-invariant filters within blocks. The separation parameters are updated from one block to the next and the separation result over a block can be used to initialize the separation of the next block. Frame-wise algorithms can be considered as particular cases of block-wise algorithms, with single-frame blocks, and hybrid methods may combine block-wise and frame-wise processing. Notice that, depending on the implementation, some of these methods can run online.

Most block-wise systems use ICA, either in the temporal domain [Anemüller 99] being limited to anechoic setups, or instantaneous mixtures [Hild 02, Aichner 03, Prieto 05], or in the STFT domain, again for instantaneous mixtures [Mukai 03, Addison 06], but also for convolutive [Nakadai 09]. The general drawback is that ICA applies only to (over)determined mixtures. Also the block-wise ICA methods should account for the source permutation problem, not only across frequency bins, as usual, but across successive time blocks.

Examples of block-wise adaptation of binary-masking or LGM-based methods are more scarce. As for binary masking, a block-wise adaptation of [Araki 07] is proposed in [Loesch 09], where source separation is performed by clustering the observation vectors in the source image space. Under the LGM model, [Simon 12] describes an online block- and frame-wise adaptation of the general LGM framework proposed in [Ozerov 12].

One important problem, common to all block-wise approaches, is the difficulty to choose the block size. Indeed, the block size must assume a good trade-off between local channel stationarity (short blocks) and sufficient data to infer relevant statistics (long blocks). The latter constraint can drastically limit the dynamics of either the sources or the sensors [Loesch 09]. Other parameters such as the step-size of the iterative update equations may also be difficult to set [Simon 12]. In general, systematic convergence towards a good separation solution using a limited amount of signal statistics remains an open issue. Another LGM approach that uses an autoregressive (AR) signal model has been seen in [Yoshioka 11].

Dynamic scenarios have been also addressed in [Markovich-Golan 10], where a beamforming method for extracting multiple moving sources is proposed. This method is applicable only to over-determined mixture. Iterative and sequential approaches for speech

enhancement in reverberant environment have been proposed in [Weinstein 94] and employ an EM framework with a form of Kalman filtering. However, only the case of a determined mixture of two sources and two microphones was addressed.

Separating underdetermined time-varying convolutive mixtures using binary masking within the LGM framework was proposed in [Higuchi 14a]. The mixing filters are considered as latent variables that follow a Gaussian distribution with mean vector depending on the direction of arrival (DOA) of the corresponding source. The DOA is modeled as a discrete latent variable taking values from a finite set of angles and following a discrete hidden Markov model (HMM). A vEM algorithm is derived to perform inference of the sources and of the DOA sequence. This approach provides interesting results but it suffers from several limitations. First, the separation quality is poor, proper to binary masking approaches. Second, the capacity of the mixing filters is limited, due to the use of a discrete temporal model to represent a continuous variable (the source TDOA).

In the present chapter, we consider time-varying mixing filters and model them as hidden random variables. In contrast to [Higuchi 14a], our model for the mixing filters is an unconstrained continuous-valued temporal model. As for the source signals we use the LGcM-with-NMF discussed in Section 1.4.2 In this chapter we aim to discover improvements in separation performance, emerging from the modeling of the time-varying channel. Thus, incorporating alternative source models, such as the NMFiG from Chapter 2 is left for future research.

We must note that an earlier reference to the incorporation of a latent Bayesian continuous model into the underlying filtering, with application to speech processing, can be found in [Gannot 03]. Two schemes were proposed, namely a dual scheme with two Kalman filters applied sequentially to the signal and to the filter (the system), and a joint scheme using the approximated unscented Kalman filter, applied jointly to the signal and to the filter. Though inspiring, those schemes were applied to single-channel speech enhancement and speech dereverberation (i.e. a unique speech signal without interfering sources), and not to MASS. In the present chapter we provide a rigorous treatment of the joint, channel and LGcM-with-NMF signal estimation, using the variational approach. The proposed method may be viewed as a generalization of [Ozerov 10] to moving sources, moving microphones, or both.

3.3 AUDIO MIXTURES WITH TIME-VARYING FILTERS

In the STFT representation of an audio mixture with (1.4), the mixing matrix relates the spatial positions of the source signals and the microphones. Working with (1.4) where $\mathbf{A}_f$ does not vary with the time, implies that the positions of sources and microphones are fixed during the recordings. Such an assumption is quite restrictive for natural audio scenes where the speakers and the sensor-set can move during the recordings. We therefore generalise (1.4) to represent scenarios where the acoustic path linking the sources with the microphones is now time-varying.

To do that, the mixing equation (1.4) naturally becomes:

$$\mathbf{x}_{f\ell} = \mathbf{A}_{f\ell}\mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}, \tag{3.1}$$

with $\mathbf{A}_{f\ell}$ being now both frequency- and time-dependent. Eq. (3.1) assumes that the acoustic channel is not varying within an individual frame, which is a reasonable assumption for a wide variety of applications. Notice here, that (3.1) is also eligible to account for various environmental changes, beyond source movement, such as opening of a window or moving of furniture.

Similar with (1.15), the conditional PDF of the mixture, given channel and sources is:

$$p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell}, \mathbf{s}_{f\ell}) = \mathcal{N}_c(\mathbf{x}_{f\ell}; \mathbf{A}_{f\ell}\mathbf{s}_{f\ell}, \mathrm{v}_f\mathbf{I}_I), \tag{3.2}$$

with v_f being a parameter to be estimated. For $\mathbf{s}_{f\ell}$ we use LGcM-with-NMF, from Section 1.4.2. We present now the model for the time-varying mixing matrix.

3.3.1 THE ACOUSTIC CHANNEL

A straightforward use of (3.2) in the framework of [Ozerov 10] is unfeasible. Indeed if we consider, as in [Ozerov 10], every $\mathbf{A}_{f\ell}$ as a (matrix of) model parameters, we end up with an enormous parameter space. To circumvent this issue, we let the mixing matrix $\mathbf{A}_{f\ell}$ to be a hidden random variable and parametrise its temporal evolution instead, with much less parameters.

To do that, we vectorise $\mathbf{A}_{f\ell}$ by vertically concatenating its J columns $\{\mathbf{a}_{j,f\ell}\}_{j=1}^{J}$ into a single vector $\mathbf{a}_{:,f\ell} \in \mathbb{C}^{IJ}$, i.e. $\mathbf{a}_{:,f\ell} = \mathrm{vec}(\mathbf{A}_{f\ell}) = [\mathbf{a}_{1,f\ell}^\top \dots \mathbf{a}_{J,f\ell}^\top]^\top$. In the following $\mathbf{a}_{:,f\ell}$ is referred to as the *mixing vector*. Then we assume that for every frequency f the sequence of the L latent mixing vectors: $\mathbf{a}_{:,f1:L}$ is ruled by a first-order LDS, where the prior distribution and the process noise are assumed complex Gaussian:

$$p(\mathbf{a}_{:,f\ell}|\mathbf{a}_{:,f\ell-1}) = \mathcal{N}_c\left(\mathbf{a}_{:,f\ell}; \mathbf{a}_{:,f\ell-1}, \mathbf{\Sigma}_f^a\right), \tag{3.3}$$

$$p(\mathbf{a}_{:,f1}) = \mathcal{N}_c\left(\mathbf{a}_{:,f1}; \boldsymbol{\mu}_f^a, \mathbf{\Sigma}_f^a\right). \tag{3.4}$$

The mean vector $\boldsymbol{\mu}_f^a \in \mathbb{C}^{IJ}$ and the *evolution* covariance matrix $\mathbf{\Sigma}_f^a \in \mathbb{C}^{IJ\times IJ}$ are model parameters to be estimated. Note, that $\mathbf{\Sigma}_f^a$ is expected to reflect the amplitude of variations in the channel. Also, (1.4) corresponds to the particular case in the proposed model when $\mathbf{\Sigma}_f^a = \mathbf{0}_{IJ\times IJ}$. Indeed, in that case the latent state $\mathbf{a}_{:,f\ell}$ collapse to $\mathbf{a}_{:,f1}$ and the mixing matrix $\mathbf{A}_{f\ell}$ reduces to its time-invariant version $\mathbf{A}_f$.

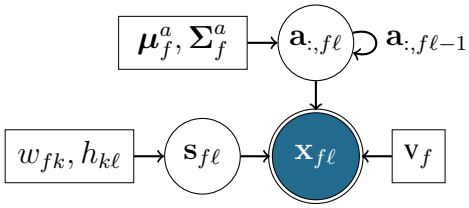

Figure 3.1: Graphical model for time-varying convolutive mixtures with NMF source model. Latent variables are represented with circles, observations with double circles, deterministic parameters with rectangles, and temporal dependencies with self loops.

3.3.2 The Complete Data Probability Distribution

The *complete data probability distribution* of all hidden variables: $\mathcal{H} = \{\mathbf{a}_{:,f\ell}, \mathbf{c}_{f\ell}, \mathbf{s}_{f\ell}\}_{f,\ell=1}^{F,L}$, observations: $\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}$, and model parameters: $\theta = \left\{\boldsymbol{\mu}_f^a, \boldsymbol{\Sigma}_f^a, w_{fk}, h_{k\ell}, \mathrm{v}_f\right\}_{f,\ell,k=1}^{F,L,K}$ writes:

$$p(\mathcal{H}, \mathbf{x}_{1:F1:L}; \theta) = \prod_{f=1}^{F} p(\mathbf{a}_{:,f1}) \prod_{\ell=2}^{L} p(\mathbf{a}_{:,f\ell}|\mathbf{a}_{:,f\ell-1}) \times \prod_{f,\ell=1}^{F,L} p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell}, \mathbf{s}_{f\ell}) \prod_{f,\ell,k=1}^{F,L,K} p(c_{k,f\ell}). \tag{3.5}$$

The complete graphical model of the proposed probabilistic model for audio source separation of time-varying convolutive mixtures can be seen in Fig. 3.1.

3.4 The vEMoVE Algorithm

Exact inference of the posterior distribution $p(\mathcal{H}|\mathbf{x}_{1:F1:L}; \theta)$ is intractable[1] for (3.5). Therefore, we construct a vEM to infer the sources, the mixing matrices and estimate the model parameters. We call the proposed algorithm *variational EM for moving environments* (vEMoVE).

In the logic of Section 1.4.1, we approximate the posterior as $q(\mathcal{H}) \approx p(\mathcal{H}|\mathbf{x}_{1:F1:L}; \theta)$ with:

$$q(\mathcal{H}) = \prod_{f=1}^{F} q(\mathbf{a}_{:,f1:L}) \prod_{f,\ell=1}^{F,L} q(\mathbf{c}_{f\ell}). \tag{3.6}$$

Each factor of $q(\mathcal{H})$ is computed with (1.8). At the E step of the vEMoVE, we first compute $q(\mathbf{c}_{f\ell})$ having at hand a previous estimate for $q(\mathbf{a}_{:,f\ell})$, and then compute $q(\mathbf{a}_{:,f\ell})$ using the just computed $q(\mathbf{c}_{f\ell})$. In the M step update θ by maximising $\mathcal{L}(\theta)$ with (1.9).

[1] The PDF of the random variable $\mathbf{A}_{f\ell}\mathbf{s}_{f\ell}$, that is a product of two Gaussian r.v.'s, is intractable.

3.4.1 E STEP

For clarity we express the E step as three substeps. The E-$\mathbf{a}_{:,f\ell}$ step computes $q(\mathbf{a}_{:,f\ell})$. The E-$\mathbf{c}_{f\ell}$ step computes $q(\mathbf{c}_{f\ell})$. And the E-$\mathbf{s}_{f\ell}$ step that computes $q(\mathbf{s}_{f\ell})$.

E-$\mathbf{a}_{:,f\ell}$ step With (1.8) it is straightforward to show that the joint posterior distribution of the mixing vector sequence writes:

$$q(\mathbf{a}_{:,f1:L}) \propto p(\mathbf{a}_{:,f1}) \prod_{\ell=2}^{L} p(\mathbf{a}_{:,f\ell}|\mathbf{a}_{:,f\ell-1}) \prod_{\ell=1}^{L} \exp\left(\mathbb{E}_{q(\mathbf{s}_{f\ell})}\left[\log p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell}, \mathbf{s}_{f\ell})\right]\right). \quad (3.7)$$

Analysing the expectation in (3.7), we have:

$$\exp\left(\mathbb{E}_{q(\mathbf{s}_{f\ell})}\left[\log p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell}, \mathbf{s}_{f\ell})\right]\right) \propto \quad (3.8)$$

$$\exp\left(-\frac{1}{\mathrm{v}_f}\mathrm{tr}\left\{-\mathbf{x}_{f\ell}^{\mathrm{H}}\mathbf{A}_{f\ell}\hat{\mathbf{s}}_{f\ell} - \left(\mathbf{A}_{f\ell}\hat{\mathbf{s}}_{f\ell}\right)^{\mathrm{H}}\mathbf{x}_{f\ell} + \mathbf{A}_{f\ell}^{\mathrm{H}}\mathbf{Q}_{f\ell}^{\eta s}\mathbf{A}_{f\ell}\right\}\right) = \quad (3.9)$$

$$\mathcal{N}_c(\mathbf{a}_{:,f\ell}; \boldsymbol{\mu}_{f\ell}^{\iota a}, \boldsymbol{\Sigma}_{f\ell}^{\iota a}), \quad (3.10)$$

with $\hat{\mathbf{s}}_{f\ell} = \mathbb{E}_{q(\mathbf{s}_{f\ell})}[\mathbf{s}_{f\ell}]$, $\mathbf{Q}_{f\ell}^{\eta s} = \mathbb{E}_{q(\mathbf{s}_{f\ell})}[\mathbf{s}_{f\ell}\mathbf{s}_{f\ell}^{\mathrm{H}}]$ computed at the E-$\mathbf{s}_{f\ell}$ step. And where:

$$\boldsymbol{\Sigma}_{f\ell}^{\iota a} = \left(\mathbf{Q}_{f\ell}^{\eta s\top} \otimes \frac{\mathbf{I}_I}{\mathrm{v}_f}\right)^{-1}, \quad (3.11)$$

$$\boldsymbol{\mu}_{f\ell}^{\iota a} = \boldsymbol{\Sigma}_{f\ell}^{\iota a}\mathrm{vec}\left(\frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}\hat{\mathbf{s}}_{f\ell}^{\mathrm{H}}\right). \quad (3.12)$$

$\mathcal{N}_c(\boldsymbol{\mu}_{f\ell}^{\iota a}; \mathbf{a}_{:,f\ell}, \boldsymbol{\Sigma}_{f\ell}^{\iota a})$ can be seen as an *observation PDF* of $\boldsymbol{\mu}_{f\ell}^{\iota a}$, given the hidden *state* $\mathbf{a}_{:,f\ell}$.

In the vEM we need the posterior distribution $q(\mathbf{a}_{:,f\ell})$, for all frames ℓ. To calculate $q(\mathbf{a}_{:,f\ell})$ we use the *Kalman smoother* algorithm [Bishop 06]; a recursive algorithm that consists of a forward pass and a backward pass. The two passes are afterwards combined to give $q(\mathbf{a}_{:,f\ell})$:

$$q(\mathbf{a}_{:,f\ell}) = \mathcal{N}_c\left(\mathbf{a}_{:,f\ell}; \hat{\mathbf{a}}_{:,f\ell}, \boldsymbol{\Sigma}_{f\ell}^{\eta a}\right), \quad (3.13)$$

with covariance matrix $\boldsymbol{\Sigma}_{f\ell}^{\eta a} \in \mathbb{C}^{IJ\times IJ}$ and mean vector $\hat{\mathbf{a}}_{:,f\ell} \in \mathbb{C}^{IJ}$, given with:

$$\boldsymbol{\Sigma}_{f\ell}^{\eta a} = \left(\boldsymbol{\Sigma}_{f\ell}^{\phi a\,-1} + \boldsymbol{\Sigma}_{f\ell}^{\beta a\,-1}\right)^{-1}, \quad (3.14)$$

$$\hat{\mathbf{a}}_{:,f\ell} = \boldsymbol{\Sigma}_{f\ell}^{\eta a}\left(\boldsymbol{\Sigma}_{f\ell}^{\phi a\,-1}\boldsymbol{\mu}_{f\ell}^{\phi a} + \boldsymbol{\Sigma}_{f\ell}^{\beta a\,-1}\boldsymbol{\mu}_{f\ell}^{\beta a}\right), \quad (3.15)$$

with $\boldsymbol{\Sigma}_{f\ell}^{\phi a}, \boldsymbol{\mu}_{f\ell}^{\phi a}$ provided by the forward pass, and with $\boldsymbol{\Sigma}_{f\ell}^{\beta a}, \boldsymbol{\mu}_{f\ell}^{\beta a}$ provided by the backward pass. We now detail the forward and backward passes.

E-$\mathbf{a}_{:,f\ell}$ step - (forward pass) The forward pass recursively provides the joint distribution of $\mathbf{a}_{:,f\ell}$ and the causal observations. The mean vector $\boldsymbol{\mu}_{f\ell}^{\phi a} \in \mathbb{C}^{IJ}$ and covariance matrix $\boldsymbol{\Sigma}_{f\ell}^{\phi a} \in \mathbb{C}^{IJ\times IJ}$ of this distribution are calculated as:

$$\boldsymbol{\Sigma}_{f\ell}^{\phi a} = \left(\boldsymbol{\Sigma}_{f\ell}^{\iota a\,-1} + \left(\boldsymbol{\Sigma}_{f\ell-1}^{\phi a} + \boldsymbol{\Sigma}_{f}^{a}\right)^{-1}\right)^{-1}, \tag{3.16}$$

$$\boldsymbol{\mu}_{f\ell}^{\phi a} = \boldsymbol{\Sigma}_{f\ell}^{\phi a}\left(\boldsymbol{\Sigma}_{f\ell}^{\iota a\,-1}\boldsymbol{\mu}_{f\ell}^{\iota a} + \left(\boldsymbol{\Sigma}_{f\ell-1}^{\phi a} + \boldsymbol{\Sigma}_{f}^{a}\right)^{-1}\boldsymbol{\mu}_{f\ell-1}^{\phi a}\right). \tag{3.17}$$

E-$\mathbf{a}_{:,f\ell}$ step - (backward pass) The backward pass recursively provides the distribution of the anti-causal observations given $\mathbf{a}_{:,f\ell}$. The mean vector $\boldsymbol{\mu}_{f\ell}^{\beta a}$ and covariance matrix $\boldsymbol{\Sigma}_{f\ell}^{\beta a}$ of this distribution are recursively calculated with:

$$\boldsymbol{\Sigma}_{f\ell}^{\zeta a} = \left(\boldsymbol{\Sigma}_{f\ell+1}^{\iota a\,-1} + \boldsymbol{\Sigma}_{f\ell+1}^{\beta a\,-1}\right)^{-1}, \tag{3.18}$$

$$\boldsymbol{\Sigma}_{f\ell}^{\beta a} = \boldsymbol{\Sigma}_{f}^{a} + \boldsymbol{\Sigma}_{f\ell}^{\zeta a}, \tag{3.19}$$

$$\boldsymbol{\mu}_{f\ell}^{\beta a} = \boldsymbol{\Sigma}_{f\ell}^{\zeta a}\left(\boldsymbol{\Sigma}_{f\ell+1}^{\iota a\,-1}\boldsymbol{\mu}_{f\ell+1}^{\iota a} + \boldsymbol{\Sigma}_{f\ell+1}^{\beta a\,-1}\boldsymbol{\mu}_{f\ell+1}^{\beta a}\right), \tag{3.20}$$

where $\boldsymbol{\Sigma}_{f\ell}^{\zeta a}$ is an intermediate matrix introduced to simplify expressions.

E-$\mathbf{c}_{f\ell}$ step Eq. (1.8), with $p(c_{k,f\ell})$ from (1.11) and $q(\mathbf{a}_{:,f\ell})$ from (3.13), yields:

$$q(\mathbf{c}_{f\ell}) \propto \exp\left(\mathbb{E}_{q(\mathbf{a}_{:,f\ell})}\left[\log p(\mathbf{x}_{f\ell}|\mathbf{A}_{f\ell}, \mathbf{s}_{f\ell})\right]\right)\prod_{k=1}^{K} p(c_{k,f\ell}) = \tag{3.21}$$

$$\mathcal{N}_c\left(\mathbf{c}_{f\ell}; \hat{\mathbf{c}}_{f\ell}, \boldsymbol{\Sigma}_{f\ell}^{\eta c}\right). \tag{3.22}$$

Eq. (3.22) resembles (1.17) although, now $\mathbf{A}_{f\ell}$ is a random variable and we use the expectations provided from $q(\mathbf{a}_{:,f\ell})$ in its place. The covariance matrix $\boldsymbol{\Sigma}_{f\ell}^{\eta c}$ and the mean vector $\hat{\mathbf{c}}_{f\ell}$ are now computed with:

$$\boldsymbol{\Sigma}_{f\ell}^{\eta c} = \left[\mathrm{diag}_K\left(\frac{1}{u_{k,f\ell}}\right) + \mathbf{G}^{\top}\frac{\boldsymbol{\Phi}_{f\ell}}{\mathrm{v}_f}\mathbf{G}\right]^{-1}, \tag{3.23}$$

$$\hat{\mathbf{c}}_{f\ell} = \boldsymbol{\Sigma}_{f\ell}^{\eta c}\mathbf{G}^{\top}\hat{\mathbf{A}}_{f\ell}^{\mathrm{H}}\frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}. \tag{3.24}$$

$\hat{\mathbf{A}}_{f\ell} = \mathbb{E}_{q(\mathbf{a}_{:,f\ell})}[\mathbf{A}_{f\ell}]$ is constructed from $\hat{\mathbf{a}}_{:,f\ell}$ (reversing the operation of column-wise vectorisation). And $\boldsymbol{\Phi}_{f\ell} = \mathbb{E}_{q(\mathbf{a}_{:,f\ell})}[\mathbf{A}_{f\ell}^{\mathrm{H}}\mathbf{A}_{f\ell}]$ with entries:[2]

$$\Phi_{jr,f\ell} = \mathbb{E}_{q(\mathbf{a}_{:,f\ell})}[\mathbf{a}_{j,f\ell}^{\mathrm{H}}\mathbf{a}_{r,f\ell}] = \mathrm{tr}\left\{\mathbb{E}_{q(\mathbf{a}_{:,f\ell})}[\mathbf{a}_{r,f\ell}\mathbf{a}_{j,f\ell}^{\mathrm{H}}]\right\} = \mathrm{tr}\left\{\mathbf{Q}_{rj,f\ell}^{\eta a}\right\}. \tag{3.25}$$

with $\mathbf{Q}_{f\ell}^{\eta a} = \mathbb{E}_{q(\mathbf{a}_{:,f\ell})}\left[\mathbf{a}_{:,f\ell}\mathbf{a}_{:,f\ell}^{\mathrm{H}}\right]$ equal:

$$\mathbf{Q}_{f\ell}^{\eta a} = \boldsymbol{\Sigma}_{f\ell}^{\eta a} + \hat{\mathbf{a}}_{:,f\ell}\hat{\mathbf{a}}_{:,f\ell}^{\mathrm{H}}, \tag{3.26}$$

[2]With the help of the cyclic property of the trace: $\mathrm{tr}\{\mathbf{AB}\} = \mathrm{tr}\{\mathbf{BA}\}$.

and $\mathbf{Q}^{\eta a}_{jr,f\ell}$ its (j,r)-th $I \times I$ sub-block.[3]

E-$\mathbf{s}_{f\ell}$ step As shown in the Appendix, $q(\mathbf{s}_{f\ell})$ is again a complex-Gaussian PDF:

$$q(\mathbf{s}_{f\ell}) = \mathcal{N}_c\left(\mathbf{s}_{f\ell}; \hat{\mathbf{s}}_{f\ell}, \mathbf{\Sigma}^{\eta s}_{f\ell}\right), \tag{3.27}$$

with parameters:

$$\mathbf{\Sigma}^{\eta s}_{f\ell} = \left[\mathrm{diag}_J\left(\frac{1}{\sum\limits_{k\in\mathcal{K}_j} u_{k,f\ell}}\right) + \frac{\mathbf{\Phi}_{f\ell}}{\mathrm{v}_f}\right]^{-1}, \tag{3.28}$$

$$\hat{\mathbf{s}}_{f\ell} = \mathbf{\Sigma}^{\eta s}_{f\ell}\hat{\mathbf{A}}^{\mathrm{H}}_{f\ell}\frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}. \tag{3.29}$$

Eq. (3.29) is structurally similar with (1.22). Notice though, that the filter-term in (1.21) is $\mathbf{A}^{\mathrm{H}}_f\mathbf{A}_f$ (a time-invariant matrix that becomes singular if $I < J$), where in (3.28) it is a full-rank frame varying matrix $\mathbf{\Phi}_{f\ell}$ that is more generic and flexible.

3.4.2 M STEP

M-$\boldsymbol{\mu}^a_f, \mathbf{\Sigma}^a_f$ step The update rules, for the LDS parameters are quite standard. The update for μ^a_f is given with, see for example Eq.(13.110) of [Bishop 06]:

$$\boldsymbol{\mu}^a_f = \hat{\mathbf{a}}_{f1}. \tag{3.30}$$

The update rule for $\mathbf{\Sigma}^a_f$ is more computationally expensive, due to the need of considering jointly two successive hidden states $\mathbf{a}_{:,f\ell}$ and $\mathbf{a}_{:,f\ell-1}$. The update writes, see for example Eq.(13.114) of [Bishop 06]:

$$\mathbf{\Sigma}^a_f = \frac{1}{L}\left(\mathbf{\Sigma}^{\eta a}_{f1} + \sum_{\ell=1}^{L-1}\left(\mathbf{Q}^{\xi a}_{11,f\ell} - \mathbf{Q}^{\xi a}_{12,f\ell} - \mathbf{Q}^{\xi a}_{21,f\ell} + \mathbf{Q}^{\xi a}_{22,f\ell}\right)\right), \tag{3.31}$$

where $\mathbf{Q}^{\xi a}_{11,f\ell}, \mathbf{Q}^{\xi a}_{12,f\ell}, \mathbf{Q}^{\xi a}_{21,f\ell}, \mathbf{Q}^{\xi a}_{22,f\ell}$ are the respective, four $IJ \times IJ$ blocks of:

$$\mathbf{Q}^{\xi a}_{f\ell} = \mathbf{\Sigma}^{\xi a}_{f\ell} + \boldsymbol{\mu}^{\xi a}_{f\ell}\boldsymbol{\mu}^{\xi a\mathrm{H}}_{f\ell}, \tag{3.32}$$

where $\mathbf{\Sigma}^{\xi a}_{f\ell}, \mu^{\xi a}_{f\ell}$ are some composite statistics:

$$\mathbf{\Sigma}^{\xi a}_{f\ell} = \begin{bmatrix} \mathbf{\Sigma}^{\zeta a^{-1}}_{f\ell} + \mathbf{\Sigma}^{a^{-1}}_f & -\mathbf{\Sigma}^{a^{-1}}_f \\ -\mathbf{\Sigma}^{a^{-1}}_f & \mathbf{\Sigma}^{\phi a^{-1}}_{f\ell} + \mathbf{\Sigma}^{a^{-1}}_f \end{bmatrix}^{-1}, \tag{3.33}$$

$$\boldsymbol{\mu}^{\xi a}_{f\ell} = \mathbf{\Sigma}^{\xi a}_{f\ell}\left[\left(\mathbf{\Sigma}^{\zeta a^{-1}}_{f\ell}\boldsymbol{\mu}^{\beta a}_{f\ell+1}\right)^\top, \left(\mathbf{\Sigma}^{\phi a^{-1}}_{f\ell}\boldsymbol{\mu}^{\phi a}_{f\ell}\right)^\top\right]^\top. \tag{3.34}$$

[3] Parcel out $\mathbf{Q}^{\eta a}_{f\ell}$ in J^2 non-overlapping $I \times I$ blocks $\left\{\mathbf{Q}^{\eta a}_{jr,f\ell}\right\}^{J,J}_{j,r=1}$

Algorithm 3. vEMoVE: A vEM for source separation of J moving sound sources.

input $\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}$, partition matrix $\mathbf{G}$, initial parameters θ.
initialize posterior statistics $\hat{\mathbf{a}}_{:,f\ell}$, $\boldsymbol{\Sigma}_{f\ell}^{\eta a}$,
initialize $\mathbf{Q}_{f\ell}^{\eta a}$ with (3.26) and then $\boldsymbol{\Phi}_{f\ell}$ with (3.25).
repeat

E step

E-$\mathbf{c}_{f\ell}$ *step:* Compute $\Sigma_{kk,f\ell}^{\eta c}$ with (3.23), $\hat{c}_{k,f\ell}$ with (3.24), then $Q_{kk,f\ell}^{\eta c}$ with (1.25).
E-$\mathbf{s}_{f\ell}$ *step:* Compute $\boldsymbol{\Sigma}_{f\ell}^{\eta s}$ with (3.28) and $\hat{\mathbf{s}}_{f\ell}$ with (3.29), then $\mathbf{Q}_{f\ell}^{\eta s}$ with (1.24).
E-$\mathbf{a}_{:,f\ell}$ *step (measurements):* Compute $\boldsymbol{\Sigma}_{f\ell}^{\iota a}$ with (3.11) and $\boldsymbol{\mu}_{f\ell}^{\iota a}$ with (3.12)
E-$\mathbf{a}_{:,f\ell}$ *step (forward pass):*
Set $\boldsymbol{\Sigma}_{f1}^{\phi a} = \left(\boldsymbol{\Sigma}_{f1}^{\iota a\,-1} + \boldsymbol{\Sigma}_{f}^{a\,-1}\right)^{-1}$ and $\boldsymbol{\mu}_{f1}^{\phi a} = \boldsymbol{\Sigma}_{f1}^{\phi a}\left(\boldsymbol{\Sigma}_{f1}^{\iota a\,-1}\boldsymbol{\mu}_{f1}^{\iota a} + \boldsymbol{\Sigma}_{f}^{a\,-1}\boldsymbol{\mu}_{f}^{a}\right)$.
for $\ell : 2$ to L
Compute $\boldsymbol{\Sigma}_{f\ell}^{\phi a}$ with (3.16) and $\boldsymbol{\mu}_{f\ell}^{\phi a}$ with (3.17).
end
E-$\mathbf{a}_{:,f\ell}$ *step (backward pass):*
Set $\boldsymbol{\Sigma}_{fL}^{\beta a} = \boldsymbol{\Sigma}_{fL}^{\phi a}$ and $\boldsymbol{\mu}_{fL}^{\beta a} = \boldsymbol{\mu}_{fL}^{\phi a}$.
for $\ell : L-1$ to 1
Compute $\boldsymbol{\Sigma}_{f\ell}^{\zeta a}$ with (3.18).
Compute $\boldsymbol{\Sigma}_{f\ell}^{\beta a}$ with (3.19) and $\boldsymbol{\mu}_{f\ell}^{\beta a}$ with (3.20).
end
E-$\mathbf{a}_{:,f\ell}$ *step (posterior):* Compute $\boldsymbol{\Sigma}_{f\ell}^{\eta a}$ with (3.14) and $\hat{\mathbf{a}}_{:,f\ell}$ with (3.15).
Compute $\mathbf{Q}_{f\ell}^{\eta a}$ with (3.26) and then $\boldsymbol{\Phi}_{f\ell}$ with (3.25).
Compute $\boldsymbol{\Sigma}_{f\ell}^{\xi a}$ with (3.33), $\boldsymbol{\mu}_{f\ell}^{\xi a}$ with (3.34), then compute $\mathbf{Q}_{f\ell}^{\xi a}$ with (3.32).

M step

M-$\boldsymbol{\mu}_f^a, \boldsymbol{\Sigma}_f^a$ *step:* Update $\boldsymbol{\mu}_f^a$ with (3.30) and $\boldsymbol{\Sigma}_f^a$ with (3.31).
M-v_f *step:* Update v_f with (3.35).
M-NMF step: Update w_{fk} with (1.28), then $h_{k\ell}$ with (1.29).

until convergence
return the estimated source images $\left\{\hat{A}_{ji,f\ell}\hat{\mathbf{s}}_{j,f\ell}\right\}_{f,\ell=1}^{F,L}$.

M-$\mathbf{v}_f$ **step** The noise variance v_f is updated similar to (1.27). The difference with (1.27) is that the mixing matrix $\mathbf{A}_f$ was a model parameter, where now it is a latent variable.

Therefore using its posterior expectation instead, we identify the update rule for v_f:

$$\mathrm{v}_f = \frac{1}{LI} \sum_{\ell=1}^{L} \left(\mathbf{x}_{f\ell}^{\mathrm{H}} \mathbf{x}_{f\ell} - 2\mathfrak{Re} \left\{ \mathbf{x}_{f\ell}^{\mathrm{H}} \hat{\mathbf{A}}_{f\ell} \hat{\mathbf{s}}_{f\ell} \right\} + \mathrm{tr} \left\{ \mathbf{Q}_{f\ell}^{\eta s} \mathbf{\Phi}_{f\ell} \right\} \right), \tag{3.35}$$

M-$w_{fk}, h_{k\ell}$ step The update rules for $w_{fk}, h_{k\ell}$ are given with (1.28) and (1.29) respectively. As for $Q_{kk,f\ell}^{\eta c}$ it is given with (1.25), although using the vEMoVE's estimates for $\Sigma_{kk,f\ell}^{\eta c}$ with (3.23) and $\hat{c}_{k,f\ell}$ with (3.24).

3.4.3 IMPLEMENTING VEMOVE

The complete vEMoVE algorithm separating J sound sources from an I-channel, time-varying mixture, is given in Algorithm 3. We would like to discuss here some notes about the LDS that allowed us to have a numerically stable implementation.

The Kalman smoother algorithm requires $\mathbf{\Sigma}_{f1}^{\phi a}, \boldsymbol{\mu}_{f1}^{\phi a}$ to be set for the first frame, and $\mathbf{\Sigma}_{fL}^{\beta a}, \boldsymbol{\mu}_{fL}^{\beta a}$ to be set for the last frame. At each iteration we set $\mathbf{\Sigma}_{f1}^{\phi a} = \left({\mathbf{\Sigma}_{f1}^{\iota a}}^{-1} + \mathbf{\Sigma}_{f}^{a-1} \right)^{-1}$ and set $\boldsymbol{\mu}_{f1}^{\phi a} = \mathbf{\Sigma}_{f1}^{\phi a} \left({\mathbf{\Sigma}_{f1}^{\iota a}}^{-1} \boldsymbol{\mu}_{f1}^{\iota a} + \mathbf{\Sigma}_{f}^{a-1} \boldsymbol{\mu}_{f}^{a} \right)$. We experimentally found that the best separation scores are attained when we first run the forward pass, then set $\mathbf{\Sigma}_{fL}^{\beta a} = \mathbf{\Sigma}_{fL}^{\phi a}$ and $\boldsymbol{\mu}_{fL}^{\beta a} = \boldsymbol{\mu}_{fL}^{\phi a}$, then run the backward pass[4].

3.5 EXPERIMENTAL STUDY

To benchmark the vEMoVE algorithm we conducted a series of experiments with 2-channel time-varying convolutive mixtures of speech. As in Chapter 2, we use [Ozerov 10] as baseline. To account for the time-varying nature of the mixtures we run [Ozerov 10] block-wise; the mixture STFT is partitioned in $P = 4$ blocks of (consecutive) frames, and [Ozerov 10] is applied to each block. As discussed in Section 3.1, the block size must assume a good trade-off between local stationarity of mixing filters and a sufficient number of data to construct relevant statistics. We used $P = 4$, as it showed better overall performance for [Ozerov 10] for the entire range of source trajectories (source movements) that we experimented. We now discuss the simulation setup and then present our results.

3.5.1 INITIALIZING THE MODEL PARAMETERS

We follow the initialisation strategies presented in Section 2.4.1. To deal with the time-varying nature of the mixtures, we apply them block-wise.

[4]The backward distribution for L-th frame is a uniform (as there is no observation for frame $L+1$). Hence $\mathbf{\Sigma}_{fL}^{\beta a}, \boldsymbol{\mu}_{fL}^{\beta a}$ are the covariance matrix and mean vector of a uniform probability distribution on $\mathbb{C}^I$. We may set $\mathbf{\Sigma}_{fL}^{\beta a} = +\infty \mathbf{I}_{IJ}$ and manipulate (3.19) and (3.20) to obtain expressions for $\mathbf{\Sigma}_{fL-1}^{\beta a}, \boldsymbol{\mu}_{fL-1}^{\beta a}$, but such scheme had reduced separation performance and we instead chose to set $\mathbf{\Sigma}_{fL}^{\beta a} = \mathbf{\Sigma}_{fL}^{\phi a}, \boldsymbol{\mu}_{fL}^{\beta a} = \boldsymbol{\mu}_{fL}^{\phi a}$.

Semi-blind initialization of NMF and filters For the *NMF parameters* we use the semi-blind procedure from Section 2.4.1. For the *mixing filters:* (initialization of $\hat{\mathbf{a}}_{:,f\ell}$) we used two strategies. In the first strategy, called *Central-A*, for each source and each block $p \in [1, P]$ of the baseline method, the BRIR corresponding to the center of the block is selected for the initialization of the corresponding column of $\mathbf{A}_f^p$ (after applying a 512-point FFT). For vEMoVE the vectorised $\mathbf{A}_f^p$ is used as initial $\hat{\mathbf{a}}_{:,f\ell}$ for all frames of the p-th block. The second strategy, called *Ones-A*, consists of setting all entries of $\mathbf{A}_f^p$ and of $\hat{\mathbf{a}}_{:,f\ell}$ to 1, $\forall f, \ell$. Obviously, this is a blind and challenging setup. In both strategies, the vEMoVe and the baseline are initialized with the same amount of true information.

Blind initialization of NMF and filters In order to deal with the time-varying mixing setup, [Dorfan 15] is applied in a block-wise manner with $P = 4$ blocks of frames, in the same way that we ran [Ozerov 10]. For each source j, the block-wise estimate of source images (STFT), are concatenated, multiplied by their complex conjugate, averaged across channels, and supplied (as an $F \times L$ matrix) to the KL-NMF algorithm [Févotte 09] yielding the initial NMF parameters for the J sources. Those parameters are provided to both the vEMoVE and the baseline method. As for the mixing vectors we use only the *Ones-A* strategy as truly blind.

Other parameters: Remaining parameters are initialized blindly: $\mathbf{\Sigma}_{f\ell}^{\eta a} = 10^3 \mathbf{I}_{IJ}, \boldsymbol{\mu}_f^a = \hat{\mathbf{a}}_{:,f1}, \mathbf{\Sigma}_f^a = \mathbf{I}_{IJ}, \forall f, \ell$. The sensor noise variance v_f, the baseline method showed the best performance when initialized with 1% of the (L, I)-average PSD of the mixture, as suggested in [Ozerov 10]. Our method behaved best with a much higher initial value for v_f, namely $1{,}000$ times the (L, I)-average PSD of the mixture.

3.5.2 SIMULATION SETUP

Artificial mixtures (for semi-blind experiments) Similar with Section 2.4.2, we used monochannel 16 kHz signals as sources, randomly chosen from the TIMIT database [Garofolo 93]. Each source signal was convolved with BRIRs from [Hummersone 13] to produce the corresponding ground truth source image. We made mixtures of $J = 3$ and $J = 4$ sources. The J source images were added to generate the mix signal. The database of [Hummersone 13] provides BRIRs for azimuthal source-to-head angles in the range -90° to 90° with a 5° step. To simulate continuous circular movements we interpolated those BRIRs at the sample level using up-sampling, delay compensation, linear interpolation, delay restoration, and downsampling. Due to memory limitations, we truncated the original $16{,}000$-tap BRIRs to either 512 or $4{,}096$ taps[5]. Choosing two different lengths enables to reveal the effect of the narrow-band assumption, see Section 1.2.2. Note that the recorded BRIRs have vanished after $4{,}096$ samples, but not after 512 samples.

To measure the effect of speed, we designed two setups for the movement of the sources around the dummy head, shown in Fig. 3.2. In Type I mixtures, s_3 always moves from 85° to 45°, and the bounds of the trajectory of all other sources is varied

[5]Hence, reducing the effective reverberation time to an extent.

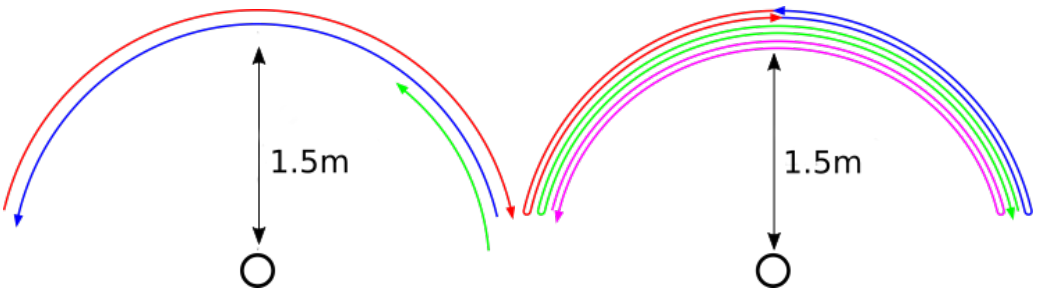

Figure 3.2: Type I (left) and II (right) source trajectories for the experiments with semi-blind initialization. In Type I, Sources s_1 (red) and s_2 (blue) move from $-\vartheta$ to ϑ and from ϑ to $-\vartheta$ respectively, Source s_3 moves from $85°$ to $45°$. In Type II, sources move: from $0°$ to $-\vartheta$ and back (s_1, red), from $0°$ to ϑ and back (s_2, blue), from $-\vartheta$ to ϑ and back (s_3, purple) and from ϑ to $-\vartheta$ and back (s_4, green); note that s_3 and s_4 move twice as fast as s_1 and s_2. In this example, $\vartheta = 75°$.

with $\vartheta \in \{15°, 30°, 45°, 60°, 75°, 90°\}$. Every trajectory is traveled within the 2s of signal duration (the signals duration is always $32{,}768$ samples), hence we had simulated different source velocities. We created four kinds of mixtures, either with filter 512 taps or 4096 taps, and with either 3 or 4 sources. The four mixtures are: *I-512-3*, *I-4096-3*, *II-512-3*,[6] and *II-512-4*. The STFT was applied to the mixed signal with a 512-sample, 50%-overlap, sine window, leading to $L = 128$ observation frames. The number of components per source was set to $|\mathcal{K}_j| = 25$. The correct number of sources in the mixture (3 or 4) is provided to the algorithms in all experiments. The number of iterations for all methods was set to 100.

Artificial mixtures (for blind experiments) For the blind experiments we create an underdetermined stereo setup of $J = 3$ simulated moving speakers from TIMIT (two male and one female). Since the blind initialization method relies on a free-field direct-path propagation model, we substitute the BRIRs with the room impulse response (RIR) simulator of AudioLabs Erlangen[7], based on the image method [Allen 79]. We defined a 2-microphone set-up with omnidirectional microphones, spaced by $d = 50$ cm. The simulated room had the same size as the one made with BRIRs. On the semi-blind experiments, we simulated sources trajectories that were crossing multiple times, to test the proposed method in a really difficult scenario. However, the binary-mask initialization method, due to being applied on blocks of time-frames, it may be subject to source permutation across blocks.[8] To avoid this problem, we simulated a new setup where the trajectories of the $J = 3$ sources are now not crossing: The 3 sources are all moving in circle of $\vartheta = 60°$ in 2 s, from $-65°$ to $-5°$ for s_1, from $-30°$ to $30°$ for s_2 and from $5°$ to $65°$ for s_3, at about 1.5 m apart from the microphone pair center (see Fig. 3.3-left).

[6] In this case we discarded the fourth source (green line, right plot, Fig. 3.2).

[7] available at www.audiolabs-erlangen.de/fau/professor/habets/software/rir-generator.

[8] Note however that [Dorfan 15] is not subject to source permutation across frequency bins since all frequencies are jointly processed in the CGMM model.

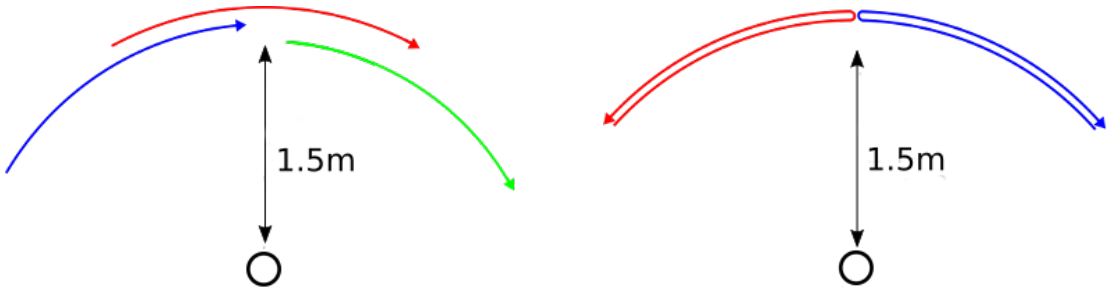

Figure 3.3: Source trajectories for the experiments with blind initialization: Simulations (left) and real recordings (right).

We simulated two reverberation times, namely $T_{60} = 680$ ms (same as in the semi-blind setup) and $T_{60} = 270$ ms (the corresponding mixtures are denoted respectively as *Mix-680* and *Mix-270*). We tested each mixture as is (noiseless case) and when corrupted with additive white Gaussian noise at SNR$= 4$ dB. This resulted in 4 configurations. All reported measures are average results over 10 mixtures using different speech signals from TIMIT and noise realization.

Real recordings Real-recordings are made in a 20 m^2 reverberant room ($T_{60} \approx 500$ ms), using $I = 2$ omnidirectional microphones in free field, placed in the center of the room, and spaced by $d = 30$ cm. For the real-recordings, the blind initialization method was shown to be much less efficient to separate 3 sources, compared to the simulated experiments, but still worked very well for 2 sources. We thus limited the present experiments to $J = 2$ sources. Two speakers (one female, one male) were asked to pronounce spontaneous speech while moving on a circle at 1.5 m from the microphones, of about $45°$, two-way opposite motions, starting respectively at about $45°$ and $-45°$ (see Fig. 3.3-right). The trajectory was traveled within 2s, hence the source movement was pretty fast. The two speakers were recorded separately, and their signals were added a posteriori to make the mix, therefore we could calculate separation scores.

3.5.3 Experiments with Semi-blind initialization

We evaluate the separation performance using standard metrics from [Vincent 06]. We first discuss detailed results for the particular but representative value of $\vartheta = 75°$. Then we report the performance of the vEMoVE with respect to ϑ and generalize the discussion.

Fig. 3.4 represents the evolution of average SDR measures with the vEM iterations, for $\vartheta = 75°$, and Mix-I. Let us recall that SDR is a general indicator that balances separation performance (that is rejection of interfering sources) and signal distortion (that measures artefacts due to the model/algorithm). Each point in the figure is an average result over all 3 sources, and 10 different runs (with different source signals). The two plots at the top correspond to mix *I-512-3* and the two plots at the bottom correspond to mix *I-4096-3*.

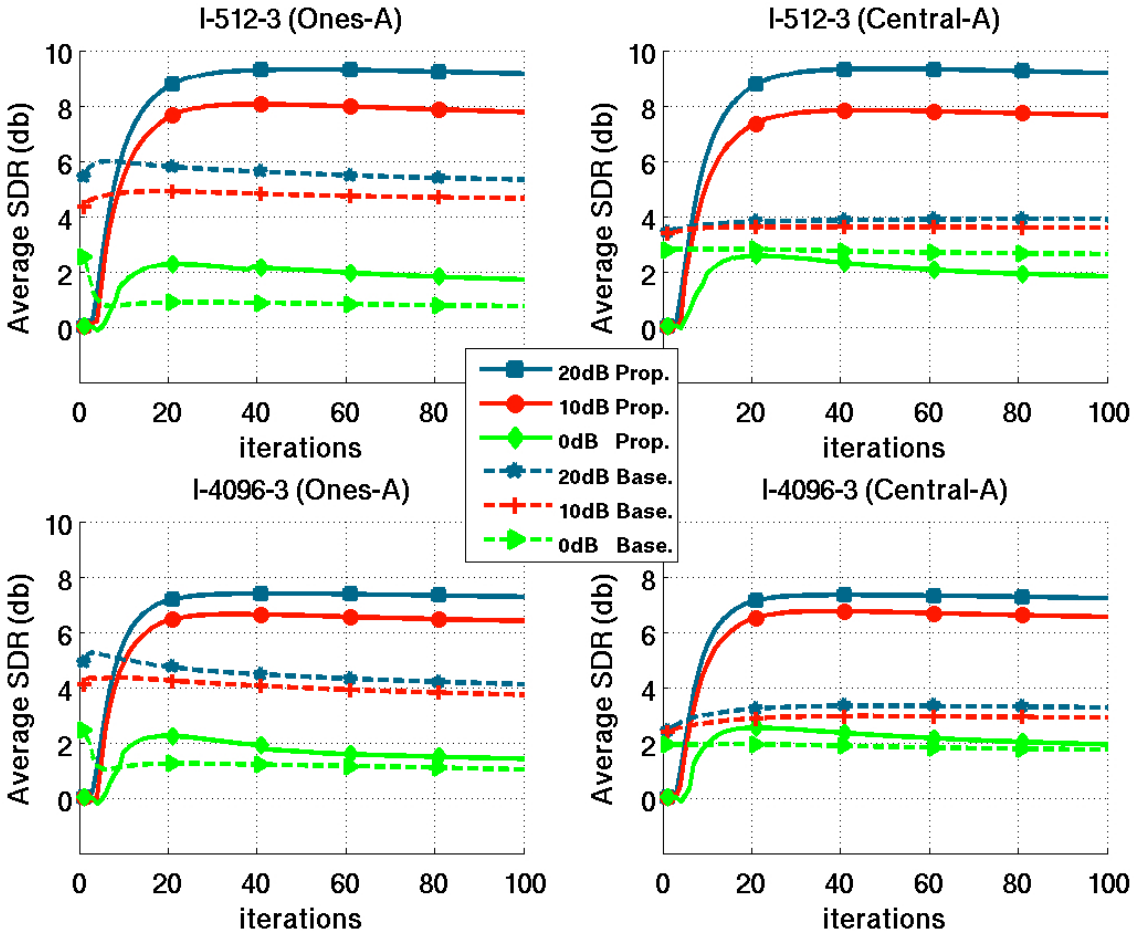

Figure 3.4: Average (over all sources) SDR vs iterations, under semi-blind initialization. (top): *I-512-3*, (bottom): *I-4096-3*, (left) column is with *Ones-A* filter initialization strategy, (right) is with *Central-A* filter initialization strategy. All experiments are at $\vartheta = 75°$.

The two plots on the left are initialised with the *Ones-A* strategy, the two on the right are initialised with *Central-A*.

Effect of NMF initialization Fig. 3.4 shows that the baseline method converges faster than the proposed method, which is natural since the baseline method operates on blocks of STFT frames and does not have the computational cost of the application of Kalman smoothing. Also, the baseline vEM has less parameters to estimate as the mixing matrix is deterministic. In *I-512-3* (*Central-A*), the proposed vEM attains SDR of ≈ 9.5 dB for $R = 20$ dB. The SDR score slightly drops to 8 dB for $R = 10$ dB, and then more abruptly decreases to 2 dB for $R = 0$ dB. SDR scores of the baseline method at $R = 20$ dB, 10 dB, and 0 dB go from 4 to 2.5 dB. The vEMoVE largely outperforms the baseline method for $R = 20$ dB and 10 dB, though in this example the baseline performs slightly better at $R = 0$ dB ($\approx +0.5$ dB over the proposed method).

Table 3.1: Average SDR and SIR for $\vartheta = 75^\circ$ with semi-blind initialization and *Ones-A*.

		SDR								SIR							
		Proposed				Baseline				Proposed				Baseline			
R	Mixture	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
20dB	I-512-3	**9.3**	**10.4**	**7.9**	–	5.5	6.5	4.0	–	**14.9**	**16.0**	**14.3**	–	10.5	12.3	8.4	–
	I-4096-3	**7.7**	**7.9**	**6.2**	–	4.7	4.6	3.0	–	**13.0**	**13.7**	**11.3**	–	10.0	9.9	6.6	–
	II-512-3	**8.4**	**8.2**	**9.5**	–	4.4	4.5	5.7	–	**13.6**	**13.8**	**16.1**	–	8.6	9.1	12.2	–
	II-512-4	**7.0**	**6.6**	**7.6**	**9.2**	3.8	3.9	4.9	5.8	**11.4**	**11.8**	**14.2**	**15.7**	7.4	8.7	9.8	11.3
10dB	I-512-3	**7.9**	**9.1**	**6.3**	–	4.8	6.0	3.1	–	**12.8**	**13.6**	**12.9**	–	9.4	11.5	7.2	–
	I-4096-3	**6.9**	**7.1**	**5.2**	–	4.2	4.4	2.5	–	**11.4**	**11.7**	**9.7**	–	9.0	9.2	5.7	–
	II-512-3	**7.1**	**6.9**	**8.2**	–	3.8	4.0	5.3	–	**11.5**	**12.2**	**13.9**	–	7.5	8.5	11.3	–
	II-512-4	**6.1**	**6.0**	**6.9**	**8.2**	3.7	3.9	4.6	5.4	**10.4**	**10.6**	**12.8**	**13.7**	6.8	8.1	8.8	10.7
0dB	I-512-3	**2.4**	**2.7**	**0.0**	–	1.1	2.3	-1.2	–	**4.3**	4.4	-0.4	–	3.7	**5.9**	**0.0**	–
	I-4096-3	**2.0**	1.9	**0.3**	–	1.8	**2.1**	-0.8	–	4.2	3.6	**-0.2**	–	**4.9**	**5.1**	-0.5	–
	II-512-3	**1.1**	**1.1**	**2.7**	–	0.0	0.4	1.7	–	**2.5**	2.1	3.9	–	2.0	**3.3**	**4.2**	–
	II-512-4	**1.8**	**1.7**	**3.4**	**3.8**	0.7	1.0	1.7	2.3	**4.2**	**3.6**	**5.3**	**5.8**	2.7	3.2	3.3	4.6

Effect of filters initialization Regarding the influence of the initialization of the mixing vectors, that is *Ones-A* vs. *Central-A*, the proposed algorithm proves to be quite robust to the filter initialisation since it attains similar results in *Ones-A* and *Central-A*. The baseline method scores lower that the proposed method for $R = 20$ dB and $R = 10$ dB, but equally well for $R = 0$ dB. Interestingly, for $R = 20$ dB and $R = 10$ dB, the baseline method scores (about 0.4–0.7 dB) higher, using the *Ones-A* (blind) configuration rather than using the *Central-A* configuration. Difficult to interpret, but a possible explanation is that we assess the performance using the source images, rather than the single-channel source signals. Although, in $R = 0$ dB the filter information delivered by *Central-A* becomes useful since now the performance of the baseline method in *Ones-A* is about 2 dB lower than that achieved with *Central-A*. In terms of SDR and for all tested R, the proposed vEM shows a clear advantage compared to the baseline method.

Effect of the narrow-band assumption As for the influence of the length of the BRIRs, we see that the performance of both proposed and baseline algorithms decreases when the BRIRs change from 512-tap to 4096-tap responses. For $R = 20$ dB and 10 dB, the decrease is of about 1.5–2 dB for the proposed method, irregardless the initialization of the mixing-vectors. The decrease is lower for the baseline method (≈ 1 dB), but this is probably related to the fact that the baseline scores are lower. For $R = 0$ dB, the influence of the BRIRs length on the performance of the proposed method is quite moderate, but this is also probably because the SDR scores are much lower than for $R = 20$ dB and 10 dB. All those manifest that (1.4) becomes a less appropriate model as the reverberation increases. Note that this is a recurrent problem in MASS in general. Our VEM is not intended to deal with this problem, but these experiments show that our VEM can provide quite remarkable SDR scores in a configuration that is *very* difficult in many aspects (underdetermined, time-varying, reverberant).

Quantitative SDR and SIR scores Table 3.1 provides per source results at iteration 100 (still averaged over 10 mixtures) and includes also SIR, for $\vartheta = 75^\circ$ and *Ones-A* filter

Table 3.2: Input SDR and SIR for the semi-blind mixtures (average over the 10 runs).

	SDR				SIR			
Mixture	s_1	s_2	s_3	s_4	s_1	s_2	s_3	s_4
I-512-3	-3.4	-1.2	-7.6	–	-2.0	-0.5	-5.9	–
I-4096-3	-2.6	-2.0	-7.5	–	-2.0	-0.5	-5.9	–
II-512-3	-5.3	-4.9	-2.1	–	-4.1	-3.7	-1.1	–
II-512-4	-7.8	-7.6	-5.3	-4.1	-6.3	-6.0	-4.1	-3.5

initialization. SIR scores focus on the ability of an MASS method to reject interfering sources. It is obvious from Table 3.1 that for $R = 20$ dB and $R = 10$ dB, the proposed vEM outperforms the baseline in both SDR and SIR for all configurations. In other words, the hierarchy discussed when analyzing Fig. 3.4 for $R = 20$ dB and $R = 10$ dB extends to per-source results, to Mix-II, and to SIR (at least for *Ones-A*). SDR improvement of the proposed method over the baseline ranges from 2.1 dB (s_2 in *II-512-4* at $R = 10$ dB) to 4.0 dB (s_1 in *II-512-3* at $R = 20$ dB). SIR improvement of the proposed method over the baseline ranges from 2.1 dB (s_2 in *I-512-3* at $R = 10$ dB) to an impressive 5.9 dB (s_3 in *I-512-3* at $R = 20$ dB). The results are particularly remarkable for the 4-source mixture configuration, with a range of output score similar to the 3-source configuration, and improvement over the baseline method up to 4.4 dB (s_3 and s_4 at $R = 20$ dB). At $R = 0$ dB the SIR results are more deteriorated for the 3-source configurations: they do not seem to indicate which method performs best (in terms of SIR). However, the SDR scores at 0 dB are all higher for the proposed method than for the baseline method, except for s_2 in mixture *I-4096-3* (only 0.2 dB below the baseline though). The improvement is however more limited than for $R = 20$ dB and $R = 10$ dB (maximum improvement is here 1.3 dB). Finally, at $R = 0$ dB, it can be noted that for the 4-source mixture, the proposed method outperforms the baseline method for all sources, and for both SDR (improvement ranges from 0.7 dB to 1.7 dB) and SIR (improvement ranges from 0.4 dB to 2 dB).

Improvement over input distortion For a source, the performance of MASS is more adequately described by the *separation gain*, that is the difference between output score and input score. Indeed, the input scores quantify how much the target source is corrupted in the mixture. A source with low input scores is more difficult to extract than a source with high input scores. In Table 3.2 we show the input SDR and input SIR scores of every source.[9] Subtracting the scores in Table 3.1 and Table 3.2, we get the SDR gains and SIR gains. We comment the results for $R = 0$ dB as the most realistic setting (remind that we are in the *Ones-A* configuration for filters). For the 3-source mixtures, vEMoVE provides a SDR gain ranging from 3.9 dB to 7.8 dB, and an SIR gain ranging from 4.1 dB to 5.8 dB. As for the 4-source mixture, the sources s_3 and s_4 score higher than s_1 and s_2 in Table 3.1, although they are moving twice as fast as s_1 and s_2 and are were expected to be

[9] We can see from Table 3.2 that the length of BRIRs does not affect the input SIR, as the entries *I-512-3* and *I-4096-3* are equal to 2nd decimal figure. For the SDR scores there is a slight degradation.

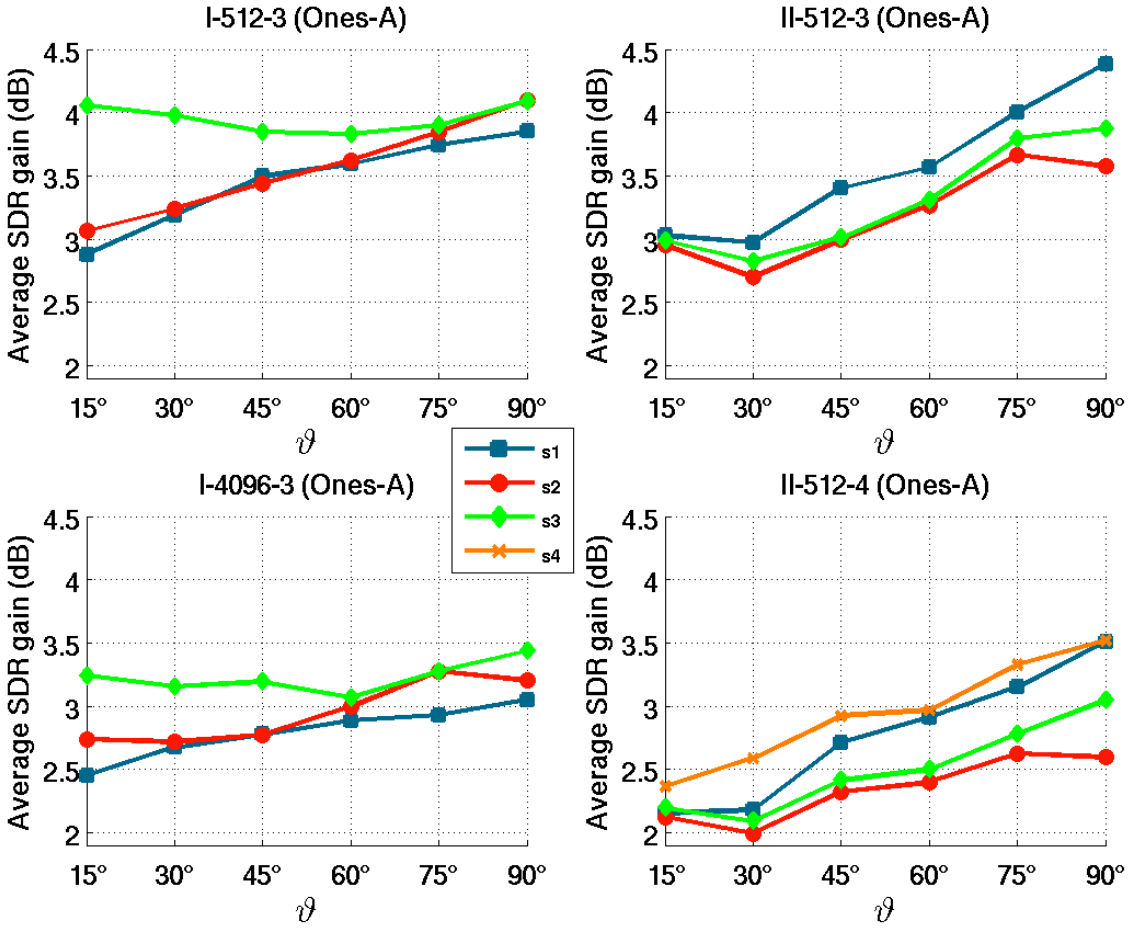

Figure 3.5: Average SDR gain for the vEM over the baseline method, with respect to Source Trajectory in semi-blind initialization, for the 4 mixture types, as a function of ϑ ($R = 20$ dB, *Ones-A* initialization).

more difficult to separate. However, they also have higher input scores, so the separation gain turns out to be quite similar overall.

Effect of speed The source's velocity of movement is proportional to ϑ. Fig. 3.5 plots the gain of the vEMoVE over the baseline method, that is the (signed) difference of the vEMoVE's SDR minus the SDR of the baseline. The results shown in Fig. 3.5 are at $R = 20\,\text{dB}$, and *Ones-A* strategy (most favorable strategy for the baseline). For *II-512-3*, we observe that except at $\vartheta = 30°$, the gain is monotonically increasing for all three sources, starting from about 3 dB at $\vartheta = 15°$ and going up to at least 3.5dB, at $\vartheta = 90°$.

There is a consistent improvement of the proposed method over the block-wise baseline, that increases with the speed of moving sources. This makes sense since the block-wise baseline method rely on the assumption that filters are stationary on each block, and this assumption gets mangled as the source speed increases. On the other hand, the pro-

Table 3.3: Average MASS scores with blind initialization (all units are in dB).

		simulated Mix-270						simulated Mix-680						real recordings		
SNR		∞			4			∞			4			N/A		
Method	Src	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR	SDR	SIR	SAR
Input	s_1	-2.3	-1.9	$+\infty$	-4.5	-1.9	4.6	-3.5	-2.9	$+\infty$	-5.5	-2.9	4.6	0.0	0.2	$+\infty$
	s_2	-3.8	-3.0	$+\infty$	-5.7	-3.0	4.6	-2.7	-1.9	$+\infty$	-4.8	-2.0	4.6	0.0	0.2	$+\infty$
	s_3	-3.1	-2.5	$+\infty$	-5.1	-2.6	4.6	-3.3	-2.7	$+\infty$	-5.3	-2.7	4.6	-	-	-
Bin-Mask	s_1	6.2	10.5	9.5	2.5	7.5	3.4	2.8	5.2	6.1	0.5	2.6	1.7	2.9	7.6	6.3
	s_2	6.2	10.8	9.4	2.0	6.9	3.4	3.8	6.9	8.2	1.2	4.7	3.1	3.1	6.4	6.6
	s_3	5.9	9.9	9.2	1.9	6.0	3.0	2.6	3.8	6.8	0.7	2.7	2.7	-	-	-
Baseline	s_1	6.0	11.1	9.7	3.2	7.9	5.3	2.3	4.9	6.4	0.7	2.6	3.4	3.5	6.7	**8.3**
	s_2	6.7	11.1	10.0	2.9	7.7	5.0	3.8	7.1	8.5	1.6	4.9	4.4	3.6	6.1	9.1
	s_3	5.9	9.7	9.5	2.8	6.7	4.8	2.5	4.4	7.1	1.1	2.8	4.2	-	-	-
Proposed	s_1	**7.5**	**13.4**	**11.5**	**5.0**	**10.0**	**8.9**	**3.3**	**6.8**	**7.8**	**1.9**	**4.0**	**6.3**	**4.2**	**7.8**	**8.3**
	s_2	**7.8**	**13.4**	**11.7**	**4.4**	**9.4**	**8.5**	**4.4**	**8.3**	**9.6**	**2.6**	**5.7**	**7.4**	**4.5**	**7.1**	**9.2**
	s_3	**7.4**	**11.7**	**11.3**	**4.6**	**7.9**	**8.5**	**3.0**	**4.9**	**8.2**	**2.3**	**3.4**	**7.3**	-	-	-

posed method seems robust to a large range of source velocity; though recall that we are in a semi-blind experimental setup. This trend is also visible on other plots. For example, for the *I-512-3* plot, the gain increases with ϑ for s_1 and s_2, from about 3 dB at $\vartheta = 15°$ to about 4 dB at $\vartheta = 90°$, whereas the gain for s_3 (whose trajectory remains independent of ϑ) is almost constant at about 4 dB. The decrease of the gain of s_3 on $\vartheta = 45°$ is attributed to the trajectories of s_1 and s_2 that interfere with s_3. Further, the curve of s_3 in *I-512-3* reveals the advantage of the proposed method even for slow movements.

3.5.4 EXPERIMENTS WITH BLIND INITIALIZATION

We report here experiments conducted with blind initialization. This series of experiments consists of two parts: the first part deals with simulated 3-speaker mixtures, and the second part deals with a 2-speaker mixture made of real recordings.

Results on artificial mixtures In Table 3.3 we report scores measured: 1) At the input mixture. 2) Using the initial estimates provided by the blind initialization method (binary masking). 3) After applying the baseline method on the mixture. 4) After applying the proposed method on the mixture. In addition to the SDR and SIR we also report the signal-to-artifacts ratios (SAR) quantifying adverse effects introduced due to the separation method. The input SDR is almost equal across sources (around -3 dB and -5 dB for the noiseless and noisy case respectively for both *Mix-270* and *Mix-680*). That indicates roughly equal power for all sources in the mix.

Let us start with the reverberant conditions *Mix-680*. At SNR $= \infty$, the average SDR (across sources) attained by the binary masking method is approximately 3 dB, hence a SDR gain of about 6 dB over the input. The corresponding SIR gain is 7.8 dB, and the output SAR[10] is about 7 dB.

[10]It make poor sense to provide SAR gain, since the source signals are intact in the mix the input SAR is $= \infty$, and applying a source separation method will lead to SAR decrease.

In *Mix-680* and SNR $= \infty$. The baseline method shows a small improvement over the binary masking scores. The proposed method shows a significant improvement, compared to any of the binary mask initialization or the baseline method. The proposed method outperforms the baseline method by: 0.5 dB to 1 dB SDR, 0.5 dB to 1.9 dB SIR, and 1.1 dB to 1.4 dB SAR. After the addition of noise (SNR $= 4$ dB), all performance measures drop significantly. For example, the average SDR for the binary masking is 2.3 dB lower than for the noiseless condition. Here, the baseline method improves the binary masking scores, by 0.3 dB SDR, 0.1 dB SIR, and 1.5 dB SAR. The proposed method outperforms the baseline method by 1.1 dB SDR, 0.9 dB SIR, and 3 dB SAR.

For *Mix-270* all methods attain higher separation scores, overall. For example, at SNR $= \infty$ the SDR of the binary masking method (averaged across sources) is 6 dB; hence an SDR-gain of about 9 dB with respect to the input. The output SIR and SAR vary from 9.2 dB to 10.8 dB (an SIR gain up to 13.8 dB). The scores (SIR measures in particular) confirm what is well-known in the literature: Binary-masking techniques show good separation performance in low-to-moderate reverberant conditions. The baseline method on the other hand exhibits comparable scores with the binary masking, slightly better on average. The vEMoVE outperforms the baseline method, by 1.4 dB in SDR, 2.2 dB in SIR, and 1.8 dB in SAR. The vEMoVE also obtains an SIR gain (with respect to the input) of 16.4 dB for Source s_2, which, we believe, is remarkable in a blind, underdetermined, dynamic (although artificial) setup. At SNR $= 4$ dB, we observe the same trend as for *Mix-680*: The baseline method improves neatly over the binary masking, and the vEMoVE significantly ameliorates over the baseline method (by 1.7 dB SDR, 1.7 dB SIR, and 3.6 dB SAR).

Results on real recordings The last three columns of Table 3.3 report the performance scores for real recording's mixture. We first notice that even if we mix two sources instead of three, the performance of the binary masking method is less notable than compared to her performance on the artificial scenarios. Evidently, separating (two) moving sources from real recordings remains a challenge, even for state-of-the-art sound processing techniques. The baseline method has an SDR improvement ≈ 0.5 dB and an SAR improvement > 2 dB, for both sources, over the binary masking. However, the baseline's SIR scores slightly degrade when compared to Binary masking. The proposed method exhibits positive gains, both over the binary-masking (initialization) and over the baseline method. The SAR scores of the proposed method are equivalent to the baseline method and notably better than the initialization. SDR improves by more than 1 dB when compared to the initialization, and by 0.7 dB to 0.9 dB when compared to the baseline method. SIR improves by 0.2 dB to 0.7 dB when compared to the initialization and by 0.7 dB to 1.1 dB when compared to the baseline method. The results demonstrate the potential application of the proposed approach in the real-world and encourage us to pursue this line of research.

3.6 CONCLUSION

In this chapter we addressed the challenging task of separating the audio sources from time-varying convolutive mixtures. We started with the time-invariant convolutive MASS framework of [Ozerov 10], where we introduced time-varying mixing filters, that were considered as hidden random variables. We modeled the mixing filters with first-order Markov chains (per frequency) with complex-Gaussian observations and transition probability distributions. Since the observations do not depend only on the filters, but also on the sources (also hidden variables), the direct application of the Kalman smoother was not possible. For this reason, we designed a vEM algorithm for source separation and parameter estimation, assuming the mixing filters and the sources to be conditionally independent given the observations (that is the mixture). An extensive evaluation campaign demonstrated the experimental advantage of the proposed vEM over two baseline methods in several speech mixtures and different initialization strategies.

It is conjectured [Girin 17] that the latent mixing filters may have higher modeling capacity than their deterministic consideration. This will justify even further our choice to model the time-varying mixing filters as hidden random variables. In the present study, the number of sources in the mixture was assumed to be known. Developing algorithms capable of counting the number of emitting sources varying over time is an open issue, and a prerequisite for a fully blind scenario. In the following chapter we address the problem of estimating and tracking the activity of the sources in a MASS framework.

CHAPTER 4

UNIFYING AUDIO SOURCE SEPARATION AND AUDIO DIARISATION

We present a statistical model for simultaneous MASS and diarisation of the audio sources in convolutive audio mixtures. The sources are modeled with LGcM-with-NMF and we introduce a temporal labeling of every source in the mixture, as active or inactive, at the STFT frame level. The labeling allows us to obtain the *state of diarisation* of the mixture. We devise an EM algorithm where the source separation process is aided by the state of diarisation, as the latter indicates the emitting sources. The state of diarisation is tracked with a Hidden Markov Model (HMM) with emission probabilities computed from the source signals. The iterative nature of the EM creates a joint treatment of the two tasks. The proposed EM is benchmarked with underdetermined 2-channel mixtures of speech; We obtain separation performance comparable with [Ozerov 10] and improve in diarisation accuracy compared to a state-of-the-art speaker diarisation pipeline.

4.1 INTRODUCTION

Speaker diarisation has emerged as an increasingly important and dedicated domain of speech researh [Anguera Miro 12]. Speaker diarisation is the problem of determining "who spoke when?". Speaker diarisation requires the unsupervised identification of the intervals during which each speaker (or generally each source) is emitting. The earliest appearance of speaker diarisation can be traced back on works on telephony data. Towards the late 1990 and early 2000 broadcast news became the main focus of research and the rise of speaker diarisation occurred for automatic annotation of television and radio transmissions [Tranter 06]. Interest in meeting recordings, practically indoor audio mixtures, grew extensively from 2002 onward [Anguera Miro 12]. Today speaker diarisation plays an important role in the analysis of meeting recordings, since it allows for such

content to be structured in speaker turns, where linguistic content and other metadata can be retrieved, as the dominant speakers, the level of interactions, emotions and so forth.

Speaker diarisation is answering to the question "who is talking, and when?" whereas MASS tries to recover the emitted signals. It is apparent the two problems are related. Since knowing the separated sources of an audio mixture, one obtains the diarisation by labeling when every source emits or is silent; On the other hand, knowing he diarisation of the sources provides, how many source are present and the relevant intervals to recover those sources.

We espy thus, that a joint formulation of MASS and diarisation can favor the performance on both sides. To this aim we propose a probabilistic formulation of MASS and audio diarisation for multichannel time-invariant convolutive mixtures.

In Section 4.2 we review the literature on joint MASS and audio diarisation. In Section 4.3 we present the proposed probabilistic model. In Section 4.4, we derive the associated EM algorithm that infers the separated source signals and the diarisation, and estimates the model parameters. In Section 4.5 we evaluate its performance in source separation against [Ozerov 10], and in speaker diarisation against [Vijayasenan 12]. In Section 4.6 we place a discussion over the materials of the chapter and future directions.

4.2 Literature Review on Joint Audio Source Separation and Diarisation

Extensive research addressing independently MASS or speaker diarisation tasks has been conducted. State of the art in MASS has been discussed in previous chapters. State-of-the-art methods on speaker and audio diarisation [Tranter 06, Anguera Miro 12] mainly consist of a pipeline starting with feature extraction from the audio mixture, typically of Mel frequency cepstral coefficients (MFCC) or spatial parameters, and proceed with speech/non-speech segmentation of the mixture and clustering of the speech segments into individual speakers, see for example [Vijayasenan 12].

Except from a series of papers by Higuchi et. al., a framework addressing jointly MASS and diarisation seems overlooked in the literature; In [Higuchi 14b, Higuchi 15] the emitting/silent state of each source is independently modeled by a factorial HMM. A simple form of LGcM-with-NMF is included to address source separation, although its restriction to a single component per source limits the representation capacity of LGcM-with-NMF, without an easy generalization. Recall, that it is empirically known [Févotte 09] that a single component (rank-1 NMF) is not enough to represent speech spectrum.

To overcome this limitation we present a probabilistic model for simultaneous diarisation and MASS of multichannel audio mixtures. We consider all possible combinations of simultaneous active sources and process their activity in a joint manner. We model the sources with the general LGcM-with-NMF framework (with rank-K NMF).

4.3 AUDIO MIXTURES WITH DIARISATION

We now present the proposed probabilistic formulation of MASS with diarisation. The new formulation can be seen as a generalization of [Ozerov 10] to include diarisation and naturally complements the models of Chapter 2 and Chapter 3.

4.3.1 THE MIXING MODEL IS AWARE OF THE DIARISATION

We want to express $\mathbf{x}_{f\ell}$ in way that encodes the activity of the sources. We have $N = 2^J$ possible configurations for the activities of the J sources; we call every configuration a *state*. We represent each state $n \in [1, N]$ as a $J \times J$ diagonal matrix $\mathbf{D}_n$ with entries:

$$D_{jj,n} = 1 \text{ if the } j^{\text{th}} \text{ source is active in state } n,$$
$$D_{jj,n} = 0 \text{ otherwise.}$$

For example, with $J = 2$, the $N = 4$ possible matrices are:

$$\mathbf{D}_1 = \begin{bmatrix} 0 & \\ & 0 \end{bmatrix}, \mathbf{D}_2 = \begin{bmatrix} 0 & \\ & 1 \end{bmatrix}, \mathbf{D}_3 = \begin{bmatrix} 1 & \\ & 0 \end{bmatrix}, \mathbf{D}_4 = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}. \tag{4.1}$$

Incorporating $\mathbf{D}_n$ in the mixing equation (1.4), we rewrite $\mathbf{x}_{f\ell}$ as:

$$\mathbf{x}_{f\ell} = \mathbf{A}_f \mathbf{D}_n \mathbf{s}_{f\ell} + \mathbf{b}_{f\ell}. \tag{4.2}$$

By choosing a state n at a time frame ℓ, we select which of the J sources comprise the mixture at the ℓ-th frame. In other words $\mathbf{D}_n$ zeroes out the inactive sources.[1]

With $Z_\ell = n, n \in [1, N]$ a categorical variable indicating the state at frame ℓ, we naturally write (see (1.15)):

$$p(\mathbf{x}_{f\ell}|\mathbf{s}_{f\ell}, Z_\ell = n) = \mathcal{N}_c\left(\mathbf{x}_{f\ell}; \mathbf{A}_f \mathbf{D}_n \mathbf{s}_{f\ell}, \mathrm{v}_f \mathbf{I}_I\right), \tag{4.3}$$

where $\mathbf{A}_f, \mathrm{v}_f$ are parameters to be estimated. As for the source $\mathbf{s}_{f\ell}$ we use LGcM-with-NMF from Section 1.4.2. We now present the novel model for the state.

4.3.2 THE STATE OF DIARISATION

The activity of each sound source varies with time, hence the state is to be estimated for every frame ℓ. The state variable Z_ℓ is modeled with an HMM:

$$p(Z_1 = n) = \lambda_n, \tag{4.4}$$
$$p(Z_\ell = n|Z_{\ell-1} = r) = T_{nr}, \tag{4.5}$$

with $\lambda_n, T_{nr} \in \mathbb{R}_+, n, r \in [1, N]$ being the prior and transition parameters to be estimated.

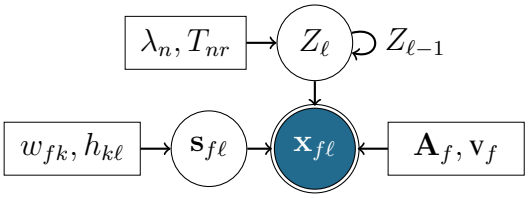

Figure 4.1: Graphical representation of our generative model for simultaneous MASS and audio diarisation. Latent variables are represented with circles, observations with double circles, deterministic parameters with rectangles, temporal dependencies with self loops.

4.3.3 THE COMPLETE DATA PROBABILITY DISTRIBUTION

In the spirit of this thesis, the complete data probability distribution of the hidden variables $\mathcal{H} = \{\mathbf{c}_{f\ell}, \mathbf{s}_{f\ell}, Z_\ell\}_{f,\ell=1}^{F,L}$, the observations $\mathbf{x}_{1:F1:L}$, and the model parameters to be estimated $\theta = \{\mathbf{A}_f, \mathrm{v}_f, w_{fk}, h_{k\ell}, T_{nr}, \lambda_n\}_{f,\ell,k,n,r=1}^{F,L,K,N,N}$ for the proposed model writes:[2]

$$p(\mathcal{H}, \mathbf{x}_{1:F1:L}; \theta) = p(Z_1) \prod_{\ell=2}^{L} p\left(Z_\ell | Z_{\ell-1}\right) \prod_{f,\ell=1}^{F,L} p(\mathbf{x}_{f\ell} | \mathbf{s}_{f\ell}, Z_\ell) \prod_{f,\ell,k=1}^{F,L,K} p(c_{k,f\ell}). \tag{4.6}$$

The graphical model of the proposed generative model for simultaneous MASS and audio diarisation is given in Fig. 4.1.

4.4 THE EMD ALGORITHM

Surprisingly the posterior probability distribution can be expressed in closed form for (4.6). This allows us to derive the EM algorithm to infer the hidden variables and estimate θ. We name our algorithm *EM for joint MASS and audio Diarisation* (EMD). We now present the E-step that computes $p(\mathcal{H}|\mathbf{x}_{1:F1:L}; \theta)$ and the M-step that updates the model parameters by maximising $\mathcal{L}(\theta)$. The complete EMD algorithm can be seen in Algorithm 4.

4.4.1 E STEP

For conciseness we describe the E-step as three sub E-steps: The E-$\mathbf{c}_{f\ell}$ step, the E-$\mathbf{s}_{f\ell}$ step and the E-Z_ℓ step. Note though here the sub E-steps are independent, whereas in a vEM they would depend on each other.

[1] In the state of "all sources are active" the $\mathbf{D}_n = \mathbf{I}_J$ and (4.2) becomes (1.4).

[2] Note that Z_ℓ is not yet evaluated to a specific n.

E-$\mathbf{c}_{f\ell}$ step We find $p(\mathbf{c}_{f\ell}, Z_\ell = n|\mathbf{x}_{1:F1:L})$ for every state. Setting $Z_\ell = n$ in (4.6):

$$p\left(\mathbf{c}_{f\ell}|Z_\ell = n, \mathbf{x}_{1:F1:L}\right) \propto p(\mathbf{x}_{f\ell}|\mathbf{s}_{f\ell}, Z_\ell = n) \prod_{k=1}^{K} p(c_{k,f\ell}) = \tag{4.7}$$

$$\mathcal{N}_c\left(\mathbf{c}_{f\ell}; \hat{\mathbf{c}}_{f\ell n}, \mathbf{\Sigma}^{\eta c}_{f\ell n}\right), \tag{4.8}$$

with mean vector $\hat{\mathbf{c}}_{f\ell n}$ and covariance matrix $\mathbf{\Sigma}^{\eta c}_{f\ell n}$ computed with:

$$\mathbf{\Sigma}^{\eta c}_{f\ell n} = \left[\text{diag}_K\left(\frac{1}{u_{k,f\ell}}\right) + \mathbf{G}^\top \mathbf{D}_n \frac{\mathbf{A}_f^{\text{H}}\mathbf{A}_f}{\mathrm{v}_f}\mathbf{D}_n\mathbf{G}\right]^{-1}, \tag{4.9}$$

$$\hat{\mathbf{c}}_{f\ell n} = \mathbf{\Sigma}^{\eta c}_{f\ell n}\mathbf{G}^\top\mathbf{D}_n\mathbf{A}_f^{\text{H}}\frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}, \tag{4.10}$$

E-$\mathbf{s}_{f\ell}$ step From the Appendix, we obtain the source posterior distribution:

$$p\left(\mathbf{s}_{f\ell}|Z_\ell = n, \mathbf{x}_{1:F1:L}\right) = \mathcal{N}_c\left(\mathbf{s}_{f\ell}; \hat{\mathbf{s}}_{f\ell n}, \mathbf{\Sigma}^{\eta s}_{f\ell n}\right), \tag{4.11}$$

with mean vector $\hat{\mathbf{s}}_{f\ell n}$ and covariance matrix $\mathbf{\Sigma}^{\eta s}_{f\ell n}$ given from:

$$\mathbf{\Sigma}^{\eta s}_{f\ell n} = \left[\text{diag}_J\left(\frac{1}{\sum\limits_{k\in\mathcal{K}_j} u_{k,f\ell}}\right) + \mathbf{D}_n\frac{\mathbf{A}_f^{\text{H}}\mathbf{A}_f}{\mathrm{v}_f}\mathbf{D}_n\right]^{-1}, \tag{4.12}$$

$$\hat{\mathbf{s}}_{f\ell n} = \mathbf{\Sigma}^{\eta s}_{f\ell n}\mathbf{D}_n\mathbf{A}_f^{\text{H}}\frac{\mathbf{x}_{f\ell}}{\mathrm{v}_f}. \tag{4.13}$$

It is interesting to see that due to the structure of (4.12), if a source is inactive at $Z_\ell = n$ (that is it has $D_{jj,n} = 0$), then also $\hat{s}_{j,f\ell n} = 0$.

E-Z_ℓ step By integrating out the $\mathbf{c}_{f\ell}$ from (4.6)[3], what remains is the posterior distribution over the state-sequence:

$$p\left(Z_{1:L}|\mathbf{x}_{1:F1:L}\right) = p(Z_1)\prod_{\ell=2}^{L} p\left(Z_\ell|Z_{\ell-1}\right)\prod_{f,\ell=1}^{F,L}\mathcal{N}_c\left(\mathbf{x}_{f\ell}; \mathbf{0}_I, \mathbf{M}_{f\ell Z_\ell}\right), \tag{4.14}$$

with the matrix $\mathbf{M}_{f\ell Z_\ell}$ for $Z_\ell = n$ computed with:

$$\mathbf{M}_{f\ell n} = \mathrm{v}_f\mathbf{I}_I + \mathbf{A}_f\mathbf{D}_n\text{diag}_J\left(\sum_{k\in\mathcal{K}_j} u_{k,f\ell}\right)\mathbf{D}_n\mathbf{A}_f^{\text{H}}. \tag{4.15}$$

[3]For the integration we use Eq. (2.115) in [Bishop 06].

Decoding the first-order HMM Eq. (4.14) is a HMM with hidden states $Z_{1:L}$, emission probability for a state $Z_\ell = n, n \in [1, N]$ given with:

$$\iota_{\ell n} = \prod_{f=1}^{F} \mathcal{N}_c \left(\mathbf{x}_{f\ell}; \mathbf{0}_I, \mathbf{M}_{f\ell n} \right), \tag{4.16}$$

and transition probability T_{nr} from state $Z_{\ell-1} = r$ to $Z_\ell = n$. We compute the posterior probability $\eta_{\ell n} = p(Z_\ell = n | \mathbf{x}_{1:F1:L})$ for every state using the well known forward-backward algorithm [Bishop 06].

In the forward-backward algorithm the posterior $\eta_{\ell n}$ is computed with:

$$\eta_{\ell n} \overset{n}{\propto} \phi_{\ell n} \beta_{\ell n}, \tag{4.17}$$

where the probabilities $\phi_{\ell n}$ and $\beta_{\ell n}$ are computed recursively:

$$\phi_{\ell n} \overset{n}{\propto} \iota_{\ell n} \sum_{r=1}^{N} T_{nr} \phi_{(\ell-1)r}, \tag{4.18}$$

$$\beta_{\ell n} \overset{n}{\propto} \sum_{r=1}^{N} T_{rn} \iota_{(\ell+1)r} \beta_{(\ell+1)r}. \tag{4.19}$$

To avoid numerical underflow, at each frame ℓ, after computing $\phi_{\ell 1:N}$ with (4.18), we normalise (by setting $\phi_{\ell n} = \phi_{\ell n} / \sum_{r=1}^{N} \phi_{\ell r}$) and proceed to the next frame. We apply the same normalisation on $\beta_{\ell n}$.

To apply the forward-backward one must set the ϕ_{1n} and β_{Ln}: At each iteration we set $\phi_{1n} = \iota_{1n} \lambda_n$ as in [Bishop 06], then run the forward recursion. Then we set $\beta_{Ln} = \phi_{Ln}$ and run the backward recursion[4].

4.4.2 M STEP

M-T_{nr}, λ_n step The update rules for the HMM parameters are quite standard (see for example Eq. (13.18), (13.19) in [Bishop 06]):

$$\lambda_n = \eta_{1n}, \tag{4.20}$$

$$T_{nr} \overset{n}{\propto} \sum_{\ell=1}^{L-1} \xi_{nr,\ell}, \tag{4.21}$$

with the *joint posterior probability of two succesive states* $\xi_{nr,\ell} = p(Z_\ell = n, Z_{\ell-1} = r | \mathbf{x}_{1:F1:L})$ that is found with, for example Eq. (13.43) in [Bishop 06]:

$$\xi_{nr,\ell} \overset{n}{\propto} \beta_{\ell n} \iota_{\ell n} T_{nr} \phi_{(\ell-1)r}. \tag{4.22}$$

[4] In theory $\beta_{Ln} = 1$ for all $n = 1 : N$, although we achieved slightly better performance in SDR by setting $\beta_{Ln} = \phi_{Ln}$.

It may happen that, for short mixtures, some transitions will not be observed with consequence the $\xi_{nr,\ell}$ for those transitions to equal zero for all the frames. Therefore, we add an artificial offset of 10^{-7} to all $\xi_{nr,\ell}$ in (4.22) prior to normalisation.[5]

M-$\mathbf{A}_f, \mathbf{v}_f$ step Consider the $\mathbf{D}_{Z_\ell}\mathbf{s}_{f\ell}$ that appears in (4.3) as a composite random variable and calculate its first and second order statistics:

$$\mathbf{o}_{f\ell} = \sum_{n=1}^{N} \eta_{\ell n} \mathbf{D}_n \hat{\mathbf{s}}_{f\ell n}, \tag{4.23}$$

$$\mathbf{Q}^{\eta o}_{f\ell} = \sum_{n=1}^{N} \eta_{\ell n} \mathbf{D}_n \left(\mathbf{\Sigma}^{\eta s}_{f\ell n} + \hat{\mathbf{s}}_{f\ell n} \hat{\mathbf{s}}^{\mathrm{H}}_{f\ell n} \right) \mathbf{D}_n. \tag{4.24}$$

The updates for $\mathbf{A}_f, \mathrm{v}_f$ are respectively:[6]

$$\mathbf{A}_f = \left(\sum_{\ell=1}^{L} \mathbf{x}_{f\ell} \mathbf{o}^{\mathrm{H}}_{f\ell} \right) \left(\sum_{\ell=1}^{L} \mathbf{Q}^{\eta o}_{f\ell} \right)^{-1}, \tag{4.25}$$

and also:

$$\mathrm{v}_f = \frac{1}{LI} \sum_{\ell=1}^{L} \left(\mathbf{x}^{\mathrm{H}}_{f\ell} \mathbf{x}_{f\ell} - 2\mathfrak{Re} \left\{ \mathbf{x}^{\mathrm{H}}_{f\ell} \mathbf{A}_f \mathbf{o}_{f\ell} \right\} + \mathrm{tr} \left\{ \mathbf{Q}^{\eta o}_{f\ell} \mathbf{A}^{\mathrm{H}}_f \mathbf{A}_f \right\} \right). \tag{4.26}$$

M-$w_{fk}, h_{k\ell}$ step The updates of $w_{fk}, h_{k\ell}$ are similar with (1.28) and (1.29) respectively, only that here $Q^{\eta c}_{kk,f\ell n}$ has to be marginalised over n:

$$w_{fk} = \frac{1}{L} \sum_{\ell,n=1}^{L,N} \eta_{\ell n} \frac{Q^{\eta c}_{kk,f\ell n}}{h_{k\ell}}, \tag{4.27}$$

$$h_{k\ell} = \frac{1}{F} \sum_{f,n=1}^{F,N} \eta_{\ell n} \frac{Q^{\eta c}_{kk,f\ell n}}{w_{fk}}, \tag{4.28}$$

with $Q^{\eta c}_{kk,f\ell n}$ the PSD of k-th component at diarisation state n:

$$Q^{\eta c}_{kk,f\ell n} = \Sigma^{\eta c}_{kk,f\ell n} + |\hat{c}_{k,f\ell n}|^2, \tag{4.29}$$

with $\Sigma^{\eta c}_{kk,f\ell n}$ given with (4.9) and $\hat{c}_{k,f\ell n}$ given with (4.10).

4.4.3 IMPLEMENTING EMD

The complete pseudo-code of the EMD algorithm can be seen in Algorithm 4.

[5] When computing probabilities of discrete events under $\overset{n}{\propto}$, those "proportional values" must be divided with their sum over n to become valid probabilities.

[6] Notice that the $\mathbf{D}_{Z_\ell}\mathbf{s}_{f\ell}$ in (4.3) plays the role of $\mathbf{s}_{f\ell}$ in (1.15).

Algorithm 4. EMD: EM for separation and diarisation of J sound sources.

input $\{\mathbf{x}_{f\ell}\}_{f,\ell=1}^{F,L}$, partition matrix $\mathbf{G}$, initial parameters θ.
construct: The 2^J matrices $\mathbf{D}_n, n \in [1, 2^J]$ with (4.1).
repeat

E step

E-$\mathbf{c}_{f\ell}$ *step:* Compute $\Sigma_{kk,f\ell}^{\eta c}$ with (4.9), $\hat{c}_{k,f\ell}$ with (4.10).
E-$\mathbf{s}_{f\ell}$ *step:* Compute $\mathbf{\Sigma}_{f\ell}^{\eta s}$ with (4.12) and $\hat{\mathbf{s}}_{f\ell}$ with (4.13).
E-Z_ℓ *step (emissions):* Compute $\iota_{\ell n}$ with (4.16).
E-Z_ℓ *step (forward pass):* Set $\phi_{1n} = \iota_{1n}\lambda_n$.
for $\ell : 2$ to L
Compute $\phi_{\ell n}$ with (4.18) and normalize it.
end
E-Z_ℓ *step (backward pass):* Set $\beta_{Ln} = \phi_{Ln}$.
for $\ell : L-1$ to 1
Compute $\beta_{\ell n}$ with (4.19) and normalize it.
end
E-Z_ℓ *step (estimate of the diarisation state):* Compute $\eta_{\ell n}$ with (4.17).
Compute $\mathbf{o}_{f\ell}$ with (4.23) and $\mathbf{Q}_{f\ell}^{\eta o}$ with (4.24).

M step

M-HMM step: Compute $\xi_{nr,\ell}$ with (4.22), normalise it, compute T_{nr} with (4.21). Compute λ_n with (4.20).
M-$\mathbf{A}_f$ *step:* Update $\mathbf{A}_f$ with (4.25).
M-v_f *step:* Update v_f with (4.26).
M-NMF step: Update w_{fk} with (4.27), then $h_{k\ell}$ with (4.28).

until convergence
return the estimated source images $\{A_{ji,f}o_{j,f\ell}\}_{f,\ell=1}^{F,L}$.

Estimation of source images and diarisation We used $\{o_{j,f\ell}\mathbf{a}_{j,f}\}_{f,\ell=1}^{F,L}$ as the STFT domain source image estimates (applying the inverse STFT with overlap-add we obtain the time domain estimates). The diarisation (classification) output $\hat{n}_\ell$ is obtained at each frame by selecting the higher value of $\eta_{\ell n}$, over n. From the corresponding $\mathbf{D}_{\hat{n}_\ell}$ we have the active sources at ℓ^{th} frame. Frames where $\eta_{\ell 1}$ is dominant are non-speech frames.

Table 4.1: Average MASS and Diarisation scores of EMD.

		Mix-8				*Mix-DC*			
		SDR	SIR	SAR	Acc.(%)	SDR	SIR	SAR	Acc.(%)
EMD	s_1	7.7	11.6	12.1	93.5	7.8	12.4	12.2	99.5
	s_2	7.9	14.9	16.6	94.3	7.3	14.0	15.1	93.2
	s_3	9.2	13.4	14.1	87.5	8.9	13.3	14.0	99.3
	avg.	8.3	13.3	14.3	91.7	8.0	13.3	13.7	97.3
Base.	s_1	7.6	12.6	12.4	89.0	7.7	12.6	12.7	87.8
	s_2	7.6	13.5	15.9	68.4	7.3	13.1	16.0	82.2
	s_3	9.0	13.1	14.8	67.4	8.8	13.0	14.8	61.8
	avg.	8.1	13.1	14.4	74.9	7.9	12.9	14.5	77.3

4.5 EXPERIMENTAL STUDY

In this section we benchmark the performance of EMD on separating and diarising underdetermined mixtures of speech.

4.5.1 SIMULATION SETUP

To assess the performance of EMD we simulated the challenging task of separating and diarising $J = 3$ sources from a synthetic convolutive stereo mixture ($I = 2$). Each source was a 27-s signal of speech, made by concatenating utterances from the TIMIT database [Garofolo 93]. Each source was made of utterances of a different person. As mixing filters, we used binaural room impulse responses (BRIRs) from [Hummersone 13] with $\text{RT}_{60} \approx 0.68$s. The three sources were positioned at azimuths $-85°, -20°, 60°$. We generated two types of mixtures: *Mix-DC* where all sources are emitting continuously. *Mix-8* where each source has balanced portions of emission and of silence so that all $N = 8$ states appear.

Baseline methods We used [Ozerov 10] for source separation and [Vijayasenan 12] for speaker diarisation. Both baselines were provided with the true number of sources. Because [Vijayasenan 12] is designed for audio streams without simultaneously emitting talkers, we considered $2^J - 1$ virtual speakers.[7] The output of [Vijayasenan 12] is a clustering of the time frames to virtual speakers. We now have to associate the virtual speakers with source combinations. A posteriori, we evaluate all possible associations of the output of [Vijayasenan 12] and the ground-truth, and report the one that gives the highest accuracy, hence favoring the baseline to a certain extent. Note that we apply a median filter (length 10 frames) on the labeling output of each source to remove any "spikes" (that is spurious activity on an isolated frame) to both EMD and [Vijayasenan 12].

[7] There are $N - 1$ virtual speakers, because [Vijayasenan 12] has a speech/non-speech detection module. The "virtual speaker" corresponding to silence is pre-detected.

MASS and Diarisation evaluation MASS performance is assessed with the SDR, SIR and SAR measures (in dB) [Vincent 06], as in previous chapters. Diarisation is assessed with Accuracy, which is defined as the percentage of frames for which a source was correctly identified (as either active if actually emitting, or inactive if actually silent).

Initializing the Model Parameters For the EMD and the MASS baseline [Ozerov 10], we use the semi-blind initialization procedure from Section 2.4.1, with $R = 10$dB. As for the transition probabilities T_{nr} of EMD we initialise them randomly and also initialise $\lambda_n = 1/N$. The diarisation baseline does not require hand-set initialisation of parameters. For the STFT analysis we used a sine window with 512 taps and 50% frame overlap, leading to $L = 1697$ frames.

4.5.2 Quantitative Results

In Table 4.1 we report detailed MASS and diarisation scores. Each entry is an average score over 10 mixture realizations with different speakers. In terms of MASS, we see that EMD performs equally well with [Ozerov 10] on both *Mix-8* and *Mix-DC*. Notably, on *Mix-8* the average SDR of the EMD is 0.2dB higher (8.3dB versus 8.1dB). This is encouraging considering that the proposed method has to learn the additional parameters to solve for diarisation.

As for the performance in diarisation, the Accuracy of the proposed method is 16.8% higher than of the baseline [Vijayasenan 12] on *Mix-8* (91.7% versus 74.9%) and 20.0% on *Mix-DC* (97.3% versus 77.3%). Although the proposed EM and the baseline are quite different in nature, and our EM is initialised with some amount of ground truth information, this is a significant effect emerging from the joint modeling of the source activity and the source signal separation.

4.5.3 Qualitative Results on Speech Diarisation

We would like to discuss here the detection capabilities of the EMD from a qualitative perspective. Fig. 4.2 illustrates the diarisation achieved for a realization of *Mix-8*. We observe that the baseline method shows a large amount of falsely-detected and undetected frames, when EMD shows significantly less misdetections. This may be attributed on the controlled initialization for the NMF parameters, although it also reveals that EMD is capable of attaining a highly accurate diarisation. Nonetheless, recall that the transition probabilities were initialized randomly, and learned from the mixture. This performance shows that a unifying framework for MASS and audio diarisation can be a wise ploy.

4.6 Conclusion

In this chapter we introduced a probabilistic framework based on LGcM-with-NMF for joint separation and diarisation of audio sources, under an elegant formulation. We de-

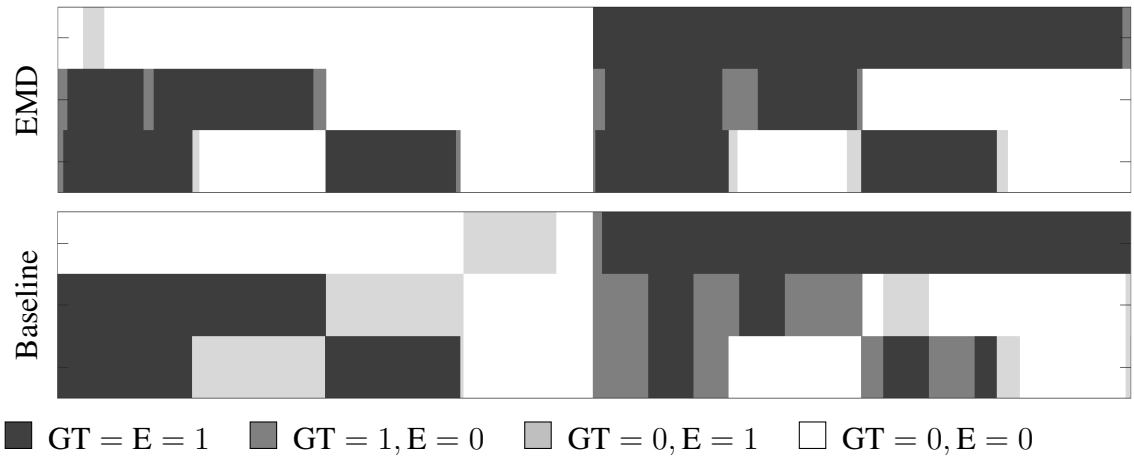

Figure 4.2: Chronogramme of diarisation. Shows the detected and undetected frames of each source's track, for the EMD and the baseline method [Vijayasenan 12], in the *Mix-8* setup. GT stands for *ground-truth*, E stands for *estimated*.

rived the associated EM algorithm for inference of the separated sources and of the diarisation. Experiments on underdetermined mixtures of speech showed competitive performance of the proposed method compared to the state of the art, in particular in diarisation scores. In the future, we would like to investigate properties that can emerge from this model as is the automatic determination of the number of sources J using $\mathbf{D}_{\hat{n}_\ell}$. Last and most important, this chapter is a not a disconnected method on its own. All previous models of this thesis can be included in a joint modular formulation to accomplish diarisation and separation of time-varying audio mixtures.

CHAPTER 5

CONCLUSION

5.1 SUMMARY AND DISCUSSION

In this thesis we studied the problem of MASS for convolutive mixtures. We made contributions in three independent and complementary directions. Our source of inspiration was [Ozerov 10], being one of the first examples of methods incorporating the LGcM-with-NMF audio signal model in a probabilistic framework for MASS.

Our journey began with a profound investigation of the role of the LGcM-with-NMF audio signal model. This search gave rise to a Bayesian alternative for LGcM-with-NMF, whose potential we demonstrated on MASS tasks.

Then, we moved to a different direction and proposed a generative model that uses LGcM-with-NMF and solves the MASS on mixtures of moving sound sources. Using the theory of *Kalman smoothing* we took care of tractability issues and the resulting method was now able to address MASS for time-varying convolutive mixtures. We tested the proposed method on underdetermined simulated and real-world mixtures of moving speakers. The experimental results revealed a significant boost in separation performance in favor for our method against a block-wise adaptation of [Ozerov 10].

Then, we envisioned and designed a generative model that jointly addresses the problem of MASS and of audio diarisation. We designed an EM algorithm, for our generative model within a framework for time-invariant audio scenes. We benchmark our EM on MASS and audio diarisation tasks against the appropriate state of the art methods; revealing promising results. Audio diarisation is significant as it can tackle the automatic estimation and tracking of the number of emitting sources in the mix.

Each of the three contributions of this thesis was presented and tested as an individual algorithm. Nonetheless, this manuscript is intended as a collection of three complementary modules enabling to construct a unified framework for simultaneous separation and diarisation of underdetermined multichannel time-varying convolutive mixtures of audio.

5.2 DIRECTIONS FOR FUTURE RESEARCH

Nowadays MASS research aims to overcome the narrowband assumption (see Section 1.2.2). To this desideratum specialized sound propagation models capable of recovering high fidelity audio signals out of highly reverberant mixtures start to appear in the MASS literature [Duong 10, Leglaive 16]. Adapting the proposals from this thesis to exploit such models is one of the natural courses for future research.

In this thesis we let aside considerations of dimensions, meaning that we did not investigate effects from the STFT analysis duration, from additional microphones, from the number of LGcM components. As also, we always provided the algorithms with the correct number of sources in the mix. Assigning LGcM components to source was considered here known in advance and fixed although, the role this assignment is a topic of active research [Ozerov 11, Bilen 16] that may reveal unknown properties of LGcM in the future. Even though, we proposed the NMFiG that appears to have an intrinsic mechanism to select how many components are actually relevant to the MASS task, hence relaxing the effect of ad-hoc setting of the number of components; for a similar mechanism for the control of NMF components see for example [Tan 13]. The diarisation enables to count and track the number of sources in the mixture, hence only the maximum number of potential sources has to be provided. Nonetheless, future research shall address in a principled way the estimation of the number of sources in the mixture; for a representative example on this direction see [Drude 14].

The major degradation of performance in LGcM based MASS appears to emerge from the initial values of the LGcM spectrum parameters. Recall that, we encountered serious difficulties over the parameter initialization for our methods especially about the source's spectrum parameters. We tackled the initialization using an additional state of the art source separation method. An extensive investigation for adequate initialisation procedures is yet to be done. However, we observed that if the source spectrum parameters were initialized with some amount of ground truth information the proposed methods were able to deliver paramount performance.

Hence, we continue to believe that probabilistic generative models enfold multifarious capabilities that may be essential in audio source separation and beyond.

PUBLICATIONS

INTERNATIONAL JOURNAL PUBLICATION

- [Kounades-Bastian 16b] Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, Sharon Gannot & Radu Horaud. *A Variational EM Algorithm for the Separation of Time-Varying Convolutive Audio Mixtures*. IEEE/ACM Transactions on Audio, Speech and Language Process. Vol.24, No.8, pp. 1408-1423, April 2016.

INTERNATIONAL CONFERENCE PUBLICATIONS

- [Kounades-Bastian 15] Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, Sharon Gannot & Radu Horaud. *A Variational EM Algorithm for the Separation of Moving Sound Sources*. In IEEE Workshop on Applications of Signal Process. to Audio and Acoustics, New Paltz, NY, Oct. 2015. *This publication received the Best Student Paper Award.*
- [Kounades-Bastian 16a] Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, Sharon Gannot & Radu Horaud. *An Inverse-Gamma Source Variance prior with factorized parametrization for audio source separation*. In IEEE Int. Conf. on Acoustics, Speech and Signal Process., Shanghai, China, March 2016.
- [Kounades-Bastian 17] Dionyssos Kounades-Bastian, Laurent Girin, Xavier Alameda-Pineda, Sharon Gannot & Radu Horaud. *An EM Algorithm for Joint Source Separation and diarisation of Multichannel Convolutive Speech Mixtures*. In IEEE Int. Conf. on Acoustics, Speech and Signal Process., New Orleans, LA, March 2017.

OTHER ARTICLE

- [Evangelidis 14] Georgios Evangelidis, Dionyssos Kounades-Bastian, Radu Horaud & Emmanouil Psarakis. *A generative model for the joint registration of multiple point sets*. European Conference on Computer Vision, Zürich, Sept. 2014.

BIBLIOGRAPHY

[Abramowitz 65] M. Abramowitz & I. Stegun. *Handbook of Mathematical Functions*. Dover Publications, 1965.

[Addison 06] W. Addison & S. Roberts. *Blind source separation with non-stationary mixing using wavelets*. In Int. Conf. Independent Component Analysis and Blind Source Separation (ICA), Charleston, SC, 2006.

[Aichner 03] R. Aichner, H. Buchner, S. Araki & S. Makino. *On-line time-domain blind source separation of nonstationary convolved signals*. In Int. Conf. Independent Component Analysis and Blind Source Separation (ICA), Nara, Japan, 2003.

[Allen 79] J. B. Allen & D. A. Berkley. *Image method for efficiently simulating small-room acoustics*. The Journal of the Acoustical Society of America, vol. 65, no. 4, pages 943–950, 1979.

[Anemüller 99] J. Anemüller & T. Gramss. *On-line blind separation of moving sound sources*. In Int. Conf. Independent Component Analysis and Blind Source Separation (ICA), Aussois, France, 1999.

[Anguera Miro 12] X. Anguera Miro, S. Bozonnet, N. Evans, C. Fredouille, G. Friedland & O. Vinyals. *Speaker Diarization: A Review of Recent Research*. IEEE Trans. Audio, Speech, and Language Process., vol. 20, no. 2, pages 356–371, 2012.

[Araki 03] S. Araki, R. Mukai, S. Makino, T. Nishikawa & H. Saruwatari. *The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech*. IEEE Trans. Speech and Audio Process., vol. 11, no. 2, pages 109–116, 2003.

[Araki 07] S. Araki, H. Sawada, R. Mukai & S. Makino. *Underdetermined blind sparse source separation for arbitrarily arranged multiple sensors*. Signal Process., vol. 87, no. 8, pages 1833–1847, 2007.

[Arberet 10] S. Arberet, A. Ozerov, N. Q. K. Duong, E. Vincent, R. Gribonval, F. Bimbot & P. Vandergheynst. *Nonnegative matrix factorization and spatial covariance model for under-determined reverberant audio source separation*. In IEEE Int. Conf. Information Sciences, Signal Process., and their Applications (ISSPA), Kuala Lumpur, Malaysia, 2010.

[Benaroya 03] L. Benaroya, L. Donagh, F. Bimbot & R. Gribonval. *Non negative sparse representation for Wiener based source separation with a single sensor*. In IEEE Int. Conf. Acoustics, Speech and Signal Process. (ICASSP), 2003.

[Bertin 10] N. Bertin, R. Badeau & E. Vincent. *Enforcing harmonicity and smoothness in Bayesian non-negative matrix factorization applied to polyphonic music transcription*. IEEE Trans. Audio, Speech, and Lang. Process., vol. 18, no. 3, pages 538–549, 2010.

[Bilen 16] C. Bilen, A. Ozerov & P. Pérez. *Automatic allocation of NTF components for user-guided audio source separation*. In IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2016.

[Bishop 06] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[Blauert 97] J. Blauert, editeur. *The psychophysics of human sound localization*. MIT Press, 1997.

[Cardoso 97] J.-F. Cardoso. *Infomax and maximum likelihood for blind source separation*. IEEE Signal Process. Letters, vol. 4, no. 4, pages 112–114, 1997.

[Cardoso 98] J. Cardoso. *Blind signal separation: Statistical principles*. Proceedings of the IEEE, vol. 9, no. 10, pages 2009–2025, 1998.

[Cemgil 09] A. T. Cemgil. *Bayesian Inference for Nonnegative matrix Factorisation Models*. Computational Intelligence and Neuroscience, 2009.

[Cherry 53] C. E. Cherry. *Some experiment the recognition of speech, withe and with two ears*. The Journal of the Acoustical Society of America, vol. 25, no. 5, pages 975–979, 1953.

[Dorfan 15] Y. Dorfan & S. Gannot. *Tree-based recursive expectation-maximization algorithm for localization of acoustic sources*. IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 23, no. 10, pages 1692–1703, 2015.

[Drude 14] L. Drude, A. Chinaev, D. H. Tran Vu & R. Haeb-Umbach. *Source counting in speech mixtures using a variational EM approach for complex Watson mixture models*. In IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Florence, Italy, 2014.

[Duong 10] N. Duong, E. Vincent & R. Gribonval. *Under-determined reverberant audio source separation using a full-rank spatial covariance model*. IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 7, pages 1830–1840, 2010.

[Ephraim 84] Y. Ephraim & D. Malah. *Speech Enhancement Using a minimum-mean square error Short-time spectral amplitude estimator*. IEEE Trans. Acoust., Speech, Signal Process., vol. 33, no. 6, pages 443–445, 1984.

[Evangelidis 14] G. Evangelidis, D. Kounades-Bastian, R. Horaud & E. Psarakis. *A generative model for the joint registration of multiple point sets*. In European Conf. Computer Vision (ECCV), 2014.

[Févotte 09] C. Févotte, N. Bertin & J.-L. Durrieu. *Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis*. Neural Computation, vol. 21, no. 3, pages 793–830, 2009.

[Gannot 01] S. Gannot, D. Burshtein & E. Weinstein. *Signal enhancement using beamforming and nonstationarity with applications to speech*. IEEE Trans. Signal Process., vol. 49, no. 8, pages 1614–1626, 2001.

[Gannot 03] S. Gannot & M. Moonen. *On the application of the unscented Kalman filter to speech processing*. In IEEE Int. Workshop Acoustic Echo and Noise Control (IWAENC), Kyoto, Japan, 2003.

[Gannot 17] S. Gannot, E. Vincent, S. Markovich-Golan & A. Ozerov. *A consolidated perspective on multi-microphone speech enhancement and source separation*. IEEE Trans. Audio, Speech, Lang. Process., 2017.

[Garofolo 93] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, N. L. Dahlgren & V. Zue. *TIMIT Acoustic-Phonetic Continuous Speech Corpus*, 1993. Linguistic Data Consortium, Philadelphia.

[Girin 17] L. Girin & R. Badeau. *On the Use of Latent Mixing Filters in Audio Source Separation*. In In 13th Int. Conf. on Latent Variable Analysis and Signal Separation (LVA/ICA), Grenoble, France, 2017.

[Higuchi 14a] T. Higuchi, N. Takamune, N. Tomohiko & H. Kameoka. *Underdetermined blind separation and tracking of moving sources based on DOA-HMM*. In IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Florence, Italy, 2014.

[Higuchi 14b] T. Higuchi, H. Takeda, N. Tomohiko & H. Kameoka. *A unified approach for underdetermined blind signal separation and source activity detection by multichannel factorial hidden Markov models*. In Interspeech, Singapore, 2014.

[Higuchi 15] T. Higuchi & H. Kameoka. *Unified approach for audio source separation with multichannel HMM and DOA mixture model*. In European Signal Process. Conf. (EUSIPCO), Nice, France, 2015.

[Hild 02] K. E. Hild, D. Erdogmus & J. C. Principe. *Blind source separation of time-varying, instantaneous mixtures using on-line algorithm*. In IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Orlando, Florida, 2002.

[Hioka 13] Y. Hioka, K. Furuya, K. Kobayashi, K. Niwa & Y. Haneda. *Underdetermined Sound Source Separation Using Power Spectrum Density Estimated by Combination of Directivity Gain*. IEEE Trans. Audio, Speech and Lang. Process., vol. 21, no. 6, 2013.

[Hjorungnes 07] A. Hjorungnes & D. Gesbert. *Complex-Valued Matrix Differentiation: Techniques and Key Results*. IEEE Trans. Signal Process., vol. 55, no. 6, pages 2740–2746, June 2007.

[Hoffman 10] M. Hoffman, D. Blei & P. Cook. *Bayesian nonparametric matrix factorization for recorded music*. In Int. Conf. Machine Learning (ICML), Haifa, Israel, 2010.

[Hummersone 13] C. Hummersone, R. Mason & T. Brookes. *A Comparison of Computational Precedence Models for Source Separation in Reverberant Environments*. J. Audio Eng. Soc., vol. 61, no. 7-8, pages 508–520, 2013.

[Hyvärinen 01] A. Hyvärinen, J. Karhunen & E. Oja, editeurs. *Independent Component Analysis*. Wiley and Sons, 2001.

[Jordan 99] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola & L. K. Saul. *An Introduction to Variational Methods for Graphical Models*. Machine Learning, vol. 37, no. 2, pages 183–233, 1999.

[Kounades-Bastian 15] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot & R. Horaud. *A Variational EM Algorithm for the Separation*

of Moving Sound Sources. In IEEE Workshop Applicat. Signal Process. to Audio and Acoust. (WASPAA), New Paltz, NY, 2015.

[Kounades-Bastian 16a] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot & R. Horaud. *An inverse-gamma source variance prior with factorized parametrization for audio source separation*. In IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2016.

[Kounades-Bastian 16b] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot & R. Horaud. *A Variational EM Algorithm for the Separation of Time-Varying Convolutive Audio Mixtures*. IEEE/ACM Trans. Audio, Speech and Language Process., vol. 24, no. 8, pages 1408–1423, 2016.

[Kounades-Bastian 17] D. Kounades-Bastian, L. Girin, X. Alameda-Pineda, S. Gannot & R. Horaud. *An EM Algorithm for Joint Source Separation and diarisation of Multichannel Convolutive Speech Mixtures*. In IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2017.

[Kowalski 10] M. Kowalski, E. Vincent & R. Gribonval. *Beyond the narrowband approximation: Wideband convex methods for underdetermined reverberant audio source separation*. IEEE/ACM Trans. Audio, Speech and Language Process., vol. 18, no. 7, pages 1818–1829, 2010.

[Lee 01] D. Lee & H. Seung. *Algorithms for non-negative matrix factorization*. Advances in Neural Information Process. Systems, vol. 13, pages 556 – 562, 2001.

[Leglaive 16] S. Leglaive, R. Badeau & G. Richard. *Multichannel Audio Source Separation with Probabilistic Reverberation Priors*. IEEE/ACM Trans. Audio, Speech and Language Process., 2016.

[Loesch 09] B. Loesch & B. Yang. *Online blind source separation based time-frequency sparseness*. In IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Taipei, Taiwan, 2009.

[Mandel 10] M. Mandel, R. J. Weiss, D. P. Ellis*et al.* *Model-based expectation-maximization source separation and localization*. IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 2, pages 382–394, 2010.

[Markovich-Golan 10] S. Markovich-Golan, S. Gannot & I. Cohen. *Subspace tracking of multiple sources and its application to speakers extraction*. In IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Dallas, TX, 2010.

[May 11] T. May, S. Van De Par & A. Kohlrausch. *A probabilistic model for robust localization based a binaural auditory front-end*. IEEE Trans. Audio, Speech, Lang. Process., vol. 19, no. 1, pages 1–13, 2011.

[Mukai 03] R. Mukai, H. Sawada, S. Araki & S. Makino. *Robust real-time blind source separation For moving speakers in a room*. In IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), 2003.

[Nakadai 09] K. Nakadai, H. Nakajima, Y. Hasegawa & H. Tsujino. *Sound source separation of moving speakers for robot audition*. In IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Taipei, Taiwan, 2009.

[Neeser 93] F. Neeser & J. Massey. *Proper complex random processes with applications to information theory*. IEEE Trans. Info. Theory, vol. 39, no. 4, pages 1293–1302, 1993.

[Ozerov 10] A. Ozerov & C. Févotte. *Multichannel Nonnegative Matrix Factorization in Convolutive Mixtures for Audio Source Separation*. IEEE Trans. Audio, Speech, Lang. Process., vol. 18, no. 3, pages 550–563, 2010.

[Ozerov 11] A. Ozerov, C. Févotte, R. Blouet & J.-L. Durrieu. *Multichannel nonnegative tensor factorization with structured constraints for user-guided audio source separation*. In Int. Conf. Acoustics, Speech and Signal Process. (ICASSP), Prague, Czech Republic, 2011.

[Ozerov 12] A. Ozerov, E. Vincent & F. Bimbot. *A General Flexible Framework for the Handling of Prior Information in Audio Source Separation*. IEEE Trans. Audio, Speech Lang. Process., vol. 20, no. 4, pages 1118–1133, 2012.

[Parra 00] L. Parra & C. Spence. *Convolutive blind separation of non-stationary sources*. IEEE Trans. Speech, Audio Process., vol. 8, no. 3, pages 320–327, 2000.

[Petersen 12] K. B. Petersen & M. S. Pedersen. *The matrix cookbook*. Version. Nov 15, 2012.

[Portnoff 80] M. Portnoff. *Time-frequency representation of digital signals and systems based short-time Fourier analysis*. IEEE Trans. Acoustics, Speech, and Signal Process., vol. 28, no. 1, pages 55–69, Feb 1980.

[Prieto 05] R. E. Prieto & P. Jinachitra. *Blind source separation for time-variant mixing systems using piecewise linear approximations*.

In IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), Philadelphia, PN, 2005.

[Sawada 04] H. Sawada, R. Mukai, S. Araki & S. Makino. *A robust and precise method for solving the permutation problem of frequency-domain blind source separation*. IEEE Trans. Speech and Audio Process., vol. 12, no. 5, pages 530–538, 2004.

[Sawada 07] H. Sawada, S. Araki, R. Mukai & S. Makino. *Grouping separated frequency components with estimating propagation model parameters in frequency-domain blind source separation.* IEEE Trans. Speech and Audio Process., vol. 15, no. 5, pages 1592–1604, 2007.

[Simon 12] L. Simon & E. Vincent. *A general framework for online audio source separation*. In Int. Conf. Latent Variable Analysis and Signal Separation (LVA/ICA), Tel-Aviv, Israel, 2012.

[Smaragdis 03] P. Smaragdis & J. Brown. *Non-negative Matrix Factorization for Polyphonic Music Transcription*. In IEEE Workshop Applicat. Signal Process. to Audio and Acoust. (WASPAA), New Paltz, NY, 2003.

[Smidl 06] V. Smidl & A. Quinn. *The Variational Bayes Method in Signal Process.* Springer-Verlag, Berlin, 2006.

[Sturmel 12] N. Sturmel, A. Liutkus, J. Pinel, L. Girin, S. Marchand, G. Richard, R. Badeau & L. Daudet. *Linear mixing models for active listening of music productions in realistic studio conditions*. In Convention of the Audio Engineering Society (AES), Budapest, Hungary, 2012.

[Tan 13] V. Y. F. Tan & C. Févotte. *Automatic Relevance Determination in Nonnegative Matrix Factorization with the beta-Divergence*. IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 35, no. 7, pages 1592–1605, 2013.

[Traa 14] J. Traa & P. Smaragdis. *Multichannel source separation and tracking with RANSAC and directional statistics*. IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 22, no. 12, pages 2233–2243, 2014.

[Tranter 06] S. Tranter & D. Reynolds. *An overview of automatic speaker diarization systems*. IEEE Trans. Audio, Speech, and Lang. Process., vol. 14, no. 5, pages 1557–1565, 2006.

[Vijayasenan 12] D. Vijayasenan, F. Valente & H. Bourlard. *Multistream speaker diarization of meetings recordings beyond MFCC and TDOA*

features. Springer handbook speech processing and speech communication, vol. 54, no. 1, 2012.

[Vincent 06] E. Vincent, R. Gribonval & C. Févotte. *Performance measurement in blind audio source separation*. IEEE Trans. Audio, Speech, Lang. Process., vol. 14, no. 4, pages 1462–1469, 2006.

[Vincent 10] E. Vincent, M. G. Jafari, S. A. Abdallah, M. D. Plumbley & M. E. Davies. *Probabilistic modeling paradigms for audio source separation*. Machine Audition: Principles, Algorithms and Systems, pages 162–185, 2010.

[Virtanen 07] T. Virtanen. *Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria*. IEEE Trans. Audio, Speech, Lang. Process., vol. 15, no. 3, pages 1066–1074, 2007.

[Virtanen 08] T. Virtanen, S. Godsill*et al. Bayesian extensions to non-negative matrix factorisation for audio signal modelling*. In IEEE Int. Conf. Acoust., Speech and Signal Process. (ICASSP), 2008.

[Wang 07] D. Wang & G. Brown, editeurs. *Computational auditory scene analysis: Principles, Algorithms and Applications*. Wiley IEEE Press, 2007.

[Weinstein 94] E. Weinstein, A. Oppenheim, M. Feder & J. Buck. *Iterative and sequential algorithms for multisensor signal enhancement*. IEEE Trans. Signal Process., vol. 42, no. 4, pages 846–859, 1994.

[Witkovsky 01] V. Witkovsky. *Computing the distribution of a linear combination of inverted gamma variables*. In Kybernetika, Prague, Czech Republic, 2001.

[Woodruff 12] J. Woodruff & D. Wang. *Binaural localization of multiple sources in reverberant and noisy environments*. IEEE Trans. Audio, Speech, Lang. Process., vol. 20, no. 5, pages 1503–1512, 2012.

[Yilmaz 04] O. Yilmaz & S. Rickard. *Blind separation of speech mixtures via time-frequency masking*. IEEE Trans. Signal Process., vol. 52, no. 7, pages 1830–1847, 2004.

[Yoshioka 11] T. Yoshioka, T. Nakatani, M. Miyoshi & H. G. Okuno. *Blind separation and dereverberation of speech mixtures by joint optimization*. IEEE Trans. Audio, Speech and Lang. Process., vol. 19, no. 1, pages 69–84, 2011.

APPENDIX

In LGcM the source components and the sources are linked with (1.14). In all algorithms derived in this thesis the posterior pdf of the components is always complex-Gaussian $\mathcal{N}_c\left(\mathbf{c};\hat{\mathbf{c}},\boldsymbol{\Sigma}^{\eta c}\right)$ with structure (omit f,ℓ,n subscripts):[1]

$$\boldsymbol{\Sigma}^{\eta c} = \left[\text{diag}_K\left(\frac{1}{u_k}\right) + \mathbf{G}^\top\boldsymbol{\Phi}\mathbf{G}\right]^{-1}, \tag{A.1}$$

$$\hat{\mathbf{c}} = \boldsymbol{\Sigma}^{\eta c}\mathbf{G}^\top\mathbf{A}^{\text{H}}\mathbf{x}. \tag{A.2}$$

Our goal is the posterior distribution of the sources, that technically is also complex-Gaussian $\mathcal{N}_c\left(\mathbf{s};\hat{\mathbf{s}},\boldsymbol{\Sigma}^{\eta s}\right)$ with parameters calculated from (1.14):

$$\boldsymbol{\Sigma}^{\eta s} = \mathbf{G}\boldsymbol{\Sigma}^{\eta c}\mathbf{G}^\top, \tag{A.3}$$

$$\hat{\mathbf{s}} = \mathbf{G}\hat{\mathbf{c}}. \tag{A.4}$$

In this Appendix we will show an efficient way to compute $\hat{\mathbf{s}}$ and $\boldsymbol{\Sigma}^{\eta s}$, without resorting to the components.

EFFICIENTLY COMPUTING THE SOURCES IN LGCM

Theorem 1 *The source posterior covariance matrix $\boldsymbol{\Sigma}^{\eta s}$ and source posterior mean vector $\hat{\mathbf{s}}$ can be computed, without resorting to the components, with:*

$$\boldsymbol{\Sigma}^{\eta s} = \left[\text{diag}_J\left(\frac{1}{\sum\limits_{k\in\mathcal{K}_j} u_k}\right) + \boldsymbol{\Phi}\right]^{-1}, \tag{A.5}$$

$$\hat{\mathbf{s}} = \boldsymbol{\Sigma}^{\eta s}\mathbf{A}^{\text{H}}\mathbf{x}. \tag{A.6}$$

[1] The structure appears in (1.17), (2.12), (3.22) and also (4.8).

Proof of Theorem 1 Apply the *Woodbury identity*[2] on (A.1), and replace in (A.3):

$$\boldsymbol{\Sigma}^{\eta s} = \mathbf{G}\boldsymbol{\Sigma}^{\eta c}\mathbf{G}^\top = \mathbf{G}\Big(\text{diag}_K\left(u_k\right) - \text{diag}_K\left(u_k\right)\mathbf{G}^\top \times$$
$$\left[\boldsymbol{\Phi}^{-1} + \mathbf{G}\text{diag}_K\left(u_k\right)\mathbf{G}^\top\right]^{-1}\mathbf{G}\text{diag}_K\left(u_k\right)\Big)\mathbf{G}^\top. \tag{A.7}$$

Observing that:

$$\mathbf{G}\text{diag}_K(u_k)\mathbf{G}^\top = \text{diag}_J\left(\sum_{r\in\mathcal{K}_j} u_r\right). \tag{A.8}$$

By replacing all four occurrences of (A.8) in (A.7), the latter becomes:

$$\boldsymbol{\Sigma}^{\eta s} = \text{diag}_J\left(\sum_{k\in\mathcal{K}_j} u_k\right) - \text{diag}_J\left(\sum_{k\in\mathcal{K}_j} u_k\right) \times$$
$$\left[\boldsymbol{\Phi}^{-1} + \text{diag}_J\left(\sum_{k\in\mathcal{K}_j} u_k\right)\right]^{-1} \text{diag}_J\left(\sum_{k\in\mathcal{K}_j} u_k\right). \tag{A.9}$$

Applying again the Woodbury identity, this time on (A.9), we have the result:

$$\boldsymbol{\Sigma}^{\eta s} = \left[\text{diag}_J\left(\frac{1}{\sum\limits_{k\in\mathcal{K}_j} u_k}\right) + \boldsymbol{\Phi}\right]^{-1}. \tag{A.10}$$

The proof is completed by substituting (A.2) in (A.4) and then identifying (A.3):

$$\hat{\mathbf{s}} = \boldsymbol{\Sigma}^{\eta s}\mathbf{A}^{\text{H}}\mathbf{x}. \tag{A.11}$$

Theorem 1 holds empirically even when $\boldsymbol{\Phi}$ is singular.

Interesting Relations *The following relations hold empirically, even if* $\boldsymbol{\Phi}$ *is singular:*

$$\hat{c}_k = \frac{u_k}{\sum\limits_{r\in\mathcal{K}_{j_k}} u_r}\hat{s}_{j_k}, \tag{A.12}$$

$$\Sigma^{\eta c}_{kk} = u_k\left(1 - \frac{u_k}{\sum\limits_{r\in\mathcal{K}_{j_k}} u_r}\left(\boldsymbol{\Phi}\boldsymbol{\Sigma}^{\eta s}\right)_{j_k j_k}\right). \tag{A.13}$$

j_k is the index of the source that k-th component belongs to, as defined with (1.14). Notice, that (A.12) is the well known Wiener filtering estimator for the LGcM components in single-channel MASS [Févotte 09].

[2] In the form $\left(\mathbf{A}^{-1} + \mathbf{G}^\top\mathbf{B}\mathbf{G}\right)^{-1} = \mathbf{A} - \mathbf{A}\mathbf{G}^\top\left(\mathbf{B}^{-1} + \mathbf{G}\mathbf{A}\mathbf{G}^\top\right)^{-1}\mathbf{G}\mathbf{A}$, see for example [Petersen 12].